Experimental evidence on the role of beliefs in prosocial behaviour

Nina Siqi Xue
Msc. Economics and Management Science
Bachelor of Commerce (Liberal Studies)

A thesis submitted for the degree of Doctor of Philosophy
at Monash University in 2023
Department of Economics

Copyright notice

Abstract

This thesis consists of three self-contained chapters exploring how the beliefs held by individuals and groups affect prosocial behaviour. Using experimental methodology, the present dissertation examines the role of beliefs in supporting self-serving behaviour and encouraging cooperation, as well as institutional factors which affect these beliefs. It advances the existing body of knowledge by providing novel insights into the interplay between beliefs about others and own behaviour, how the group decision-making environment affects how groups process new information, and finally the role of punishment mechanisms on normative beliefs and subsequent prosocial behaviour.

The introduction provides an overview of the common themes running throughout the three chapters and a brief discussion of the motivation and key findings of each chapter.

Chapter 1 focuses on beliefs, which are increasingly recognised as an important driver of behaviour, but are not straightforward to measure. We design a giving experiment to compare beliefs about others using different elicitation mechanisms when self-serving motives may compete with accuracy incentives. Consistent with a simple theoretical framework, we find evidence of a self-serving bias for non-donors when beliefs are not incentivised, while donors' beliefs are more accurate, irrespective of the elicitation mechanism. Offering a simple incentive does not reduce non-donors' underestimation of actual giving, however, a variation of the Becker-DeGroot-Marschak (BDM) procedure does appear to mitigate the negative bias in beliefs by both structuring the belief question as a question about payment and increasing the salience of monetary incentives.

Building on this, Chapter 2 is motivated by the fact that many economic decisions are made by teams, committees and boards, yet relatively little is known about how the beliefs that inform decision making in groups are formed. We conduct an experiment to examine the role of communication in belief updating. Overall, neither prior beliefs nor transfers differ between individuals and groups. Groups exhibit asymmetric updating but are not more biased than individuals. Based on text analyses of the chat data, we identify risk preferences as an important topic in communication and observe a self-serving bias for more risk-averse groups – but not for risk-averse individuals. While the group environment does not necessarily lead to more motivated beliefs, communication can amplify individual preferences and lead to more biased information processing by groups.

Chapter 3 shifts the focus to punishment mechanisms, which are commonly used to encourage cooperation, and investigates how punishment can influence beliefs about social norms. We examine whether the punisher's motives can help reconcile the conflicting results in the literature on punishment and prosocial behaviour through a novel experiment in

which the agent's outcomes are identical in two environments, but in one punishment is self-serving (i.e., benefits the punisher) while in the other it is other-regarding (i.e., benefits a third party). Self-regarding punishment reduces the social stigma of selfish behaviour, while other-regarding punishment does not. As a result, self-serving punishment is more likely to backfire.

Finally this dissertation concludes by discussing some policy implications and suggests areas for future research that follow from the contributions of this thesis.

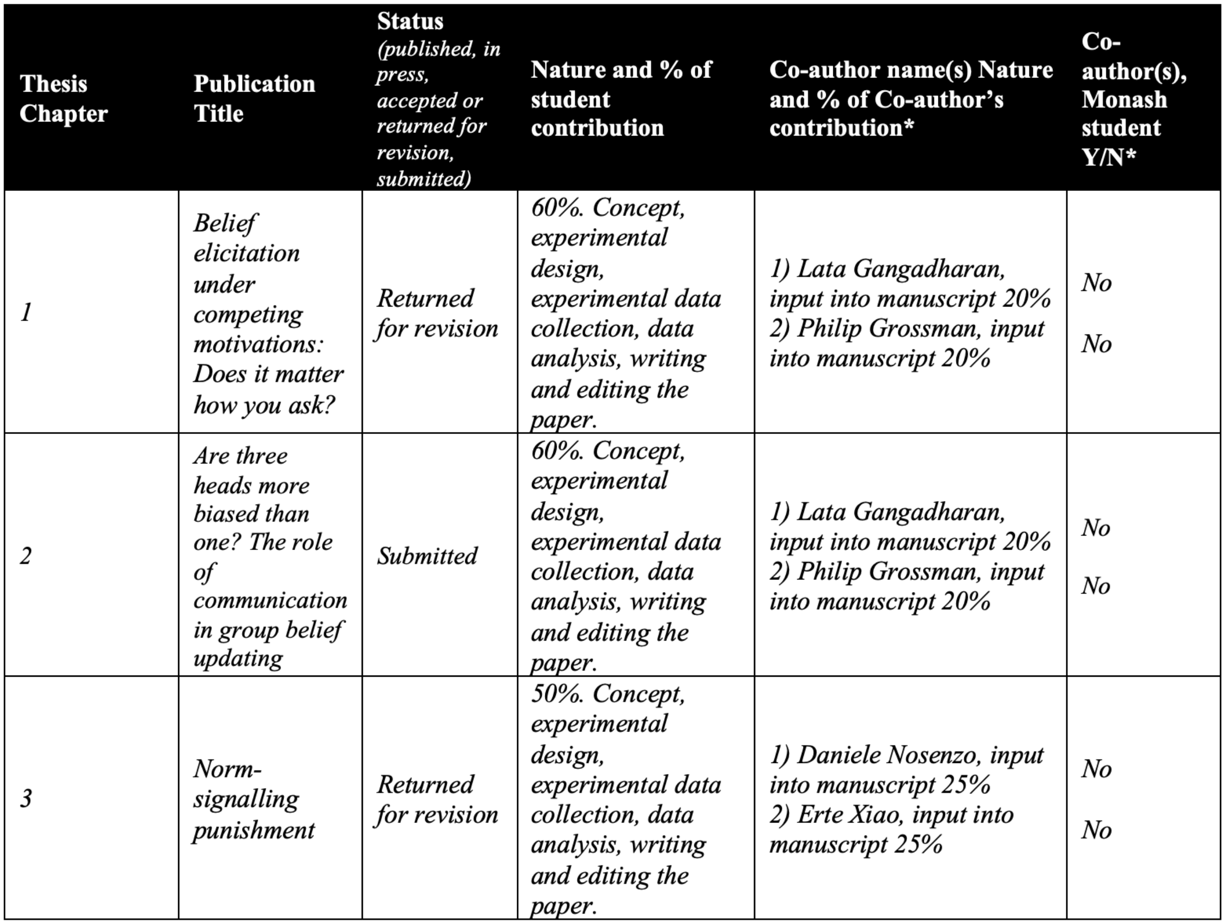
Declaration

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes zero original papers published in peer reviewed journals and two submitted publications. The core theme of the thesis is the role of beliefs in prosocial behaviour. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within The Department of Economics under the supervision of Professor Erte Xiao, Professor Lata Gangadharan and Professor Philip J. Grossman.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research. In the case of chapters 1, 2 and 3 my contribution to the work involved the following:



Thesis Chapter	**Publication Title**	**Status** *(published, in press, accepted or returned for revision, submitted)*	**Nature and % of student contribution**	**Co-author name(s) Nature and % of Co-author's contribution***	**Co-author(s), Monash student Y/N***
1	*Belief elicitation under competing motivations: Does it matter how you ask?*	*Returned for revision*	*60%. Concept, experimental design, experimental data collection, data analysis, writing and editing the paper.*	*1) Lata Gangadharan, input into manuscript 20%* *2) Philip Grossman, input into manuscript 20%*	*No* *No*
2	*Are three heads more biased than one? The role of communication in group belief updating*	*Submitted*	*60%. Concept, experimental design, experimental data collection, data analysis, writing and editing the paper.*	*1) Lata Gangadharan, input into manuscript 20%* *2) Philip Grossman, input into manuscript 20%*	*No* *No*
3	*Norm-signalling punishment*	*Returned for revision*	*50%. Concept, experimental design, experimental data collection, data analysis, writing and editing the paper.*	*1) Daniele Nosenzo, input into manuscript 25%* *2) Erte Xiao, input into manuscript 25%*	*No* *No*

I have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

I hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Acknowledgements

It wasn't always my goal to pursue a PhD, but a number of fortuitous decisions led me down this path until it became clear to me that conducting research was what I loved doing and I couldn't see myself *not* pursing a PhD in experimental economics. I am extremely grateful for the support of many people, both for helping me discover my passion and without whom this thesis would not have been possible.

To Dirk Engelmann, your lectures and open invitation to the BBE seminars sparked my interest in experimental economics. Thank you for being the first to believe in me as a researcher, I honestly would not have chosen to pursue a PhD without your kind words and support to me as a Masters student. I hope we can work together again in the future.

I could not have started my PhD journey at Monash without the support of my wonderful advisors, Lata Gangadharan, Philip Grossman and Erte Xiao. Thank you for taking a chance on me all those years ago, it's been a pleasure working with you and I sincerely hope we can continue writing papers together in the future. To Erte, thank you for your generosity with your time and advice, your honesty in your feedback and for pushing me to be a better researcher. To Lata and Phil, thank you for your warmth, mentorship and for always advocating for me. You both set the standard for not only producing high-quality research, but doing so with humility and integrity.

Many thanks also to the BET community and PhD students at Monash. Attending seminars has been one of my favourite parts of my PhD experience and I am very grateful for all the helpful feedback I have received. Special thanks to Ben – I feel very lucky to have had a teammate, co-author and friend to celebrate our successes and support each other through the rejections.

I had the pleasure of also collaborating with: Catherine Eckel, Daniele Nosenzo, Florian Schneider and Roberto Weber. Thank you for your time, patience and support. I have learnt an immense amount from working with each of you.

I am very grateful for my friends and family all around the world – thank you for putting up with me when I talk about my research and for being the best sounding boards for my ideas. To my parents, thank you for showing me the value of hard work, your endless support and for the sacrifices you made to give me a better life. Luan is so lucky to have the best Laolao and Laoye. To Valentin, thank you for moving halfway across the world to support me in pursuing my PhD, for stepping into your new role as Papa with so much enthusiasm and for giving me the time to finish my thesis. As far as rocks go, you might just be my biggest one. To Oliver, I feel so grateful to be your Mama. I sincerely hope that you too can find your passion in life.

Contents

Introduction

Alongside preferences, beliefs are central to understanding economic decision making. Beliefs play a crucial role not only in decisions involving risk, but also in social interactions in which behaviour depends on expectations about the actions of others. This dissertation consists of three self-contained chapters, each examining a different aspect of the interplay between beliefs and prosocial decision making using both laboratory and online economics experiments. First, we explore the relationship between beliefs and prosociality from the individual's perspective and examine how belief responses can vary with the elicitation method. Following this, we broaden our focus from individuals to institutions, e.g., firms, government organisations and churches, which are integral to the functioning of society. We investigate the role of the group decision-making setting on belief updating and compare this to belief updating by individuals. We also study the impact of different punishment institutions in encouraging cooperation through the channel of beliefs.

Given that beliefs are rarely observed in the "wild", experimental methods are particularly useful because we can elicit beliefs from participants rather than infer beliefs based on observed behaviour. Another major advantage of experiments is that they allow us to strip institutions down to their essential components and examine institutional features without the idiosyncrasies of particular institutions in the real world.

Beliefs are increasingly featured in economic models to better understand and predict behaviour. However, one obstacle is that beliefs are not directly observable and need to be elicited. Researchers must therefore decide which belief elicitation mechanism to use, as each method comes with advantages and disadvantages. For example, simply asking for beliefs without any incentives is the simplest approach, but may be less effective when answering the belief question requires more cognitive effort, see Charness et al. (2021) for a discussion. On the other hand, more complex incentivised mechanisms such as the binarised scoring rule (Hossain and Okui, 2013) may incentivise effort, but could be more sensitive to how the incentive is implemented (Danz et al., 2022). While previous studies have compared various elicitation methods in risky domains (e.g., Trautmann and van de Kuilen, 2015; Schotter and Trevino, 2014; Schlag et al., 2015), less is known about how the different methods compare in social domains, in particular, in settings in which beliefs may be motivated by self-serving concerns.

In **Chapter 1**, we address this gap in the literature by designing an online experiment which asks participants to make a donation to a charity and then elicits their beliefs about others' donation choices using various elicitation mechanisms. We outline a simple theoretical framework which captures the tradeoff between a desire to maximise monetary payoffs through accurate beliefs and the desire to maintain a positive self image by holding the belief that others are not generous. We find that subjects who choose not to give systematically underestimate actual giving. We observe this self-serving bias in beliefs under both a non-incentivised and a simple incentivised elicitation mechanism. However, the Karni (2009) method, a more sophisticated incentivised mechanism, reduces the bias in beliefs due to the framing of the question and the greater complexity of the method.

While examining beliefs at the individual level is a useful starting point, many important decisions are made by groups of decision makers, such as a board of directors at a company or members of a family. Groups may consist of individuals with different beliefs who must come together and agree on a common group belief. While interest in motivated beliefs is growing (e.g., Gino et al., 2016; Bénabou and Tirole, 2016; Di Tella et al., 2015), most of the literature has focused on individual beliefs. Less is known about whether biased beliefs may also be present in groups of decision makers and the role of communication in belief formation in groups. Previous literature has found that groups tend to make more selfish decisions than individuals (e.g., Bornstein and Yaniv, 1998; Kugler et al., 2007). One proposed explanation from social psychology is "schema-based distrust" (Insko and Schopler, 1987), which emphasises the important role of beliefs in potentially justifying more self-interested decisions by groups.

In **Chapter 2** we move from an individual setting to a group decision-making environment to investigate whether and how groups differ from individuals in incorporating new information into their beliefs. We conduct a laboratory experiment in which we elicit group beliefs about transfers made in a simultaneous version of the trust game and compare these to a treatment which elicits individual beliefs from group members. Our findings show that group transfers do not differ significantly from individual transfers. We are more likely to observe asymmetric updating in groups but groups are not necessarily more prone to biased beliefs. We do, however, observe a self-serving bias in more risk-averse groups, but not in risk-averse individuals. Our results suggest the potential for the group setting to exaggerate individual preferences and lead to more biased beliefs.

Another way in which institutions may influence beliefs about prosocial behaviour is through incentive mechanisms that are designed to encourage cooperation. Specifically, we focus on punishment, which has been shown to sometimes promote prosociality (e.g., Fehr and Gächter, 2002) but in other contexts backfire and result in more selfishness (e.g., Gneezy and Rustichini, 2000). While there is growing evidence on the benefits of providing normative information alongside punishment (e.g., Bicchieri et al., 2021), less is known about the specific features of punishment mechanisms that affect the communication of norms and behaviour (Bowles and Polania-Reyes, 2012). These institutional details are particularly important for the design of punishment mechanisms that aim to more effectively transmit normative information about what is appropriate and inappropriate behaviour.

In **Chapter 3**, we focus on the role of the punisher's motives in influencing beliefs about social norms and through this, the effectiveness of punishment. We design an online experiment to compare punishment that is self-serving, or intended to change the agent's behaviour for the benefit of the punisher, and punishment that is other-regarding, or intended to change the agent's behaviour for the benefit of a third party. Consistent with our theoretical prediction, we find more crowding out of prosociality from self-serving punishment due to punishment sending a weaker normative message, as compared to other-regarding punishment.

Finally, this thesis concludes by summarising key policy implications that follow from our findings, as well as interesting avenues for future research that arise as a result of this dissertation.

References

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–64.

Bicchieri, C., Dimant, E., and Xiao, E. (2021). Deviant or wrong? The effects of norm information on the efficacy of punishment. *Journal of Economic Behavior and Organization*, 188:209–235.

Bornstein, G. and Yaniv, I. (1998). Individual and group behavior in the ultimatum game: are groups more "rational" players? *Experimental Economics*, 1(1):101–108.

Bowles, S. and Polania-Reyes, S. (2012). Economic incentives and social preferences: substitutes or complements? *Journal of Economic Literature*, 50(2):368–425.

Charness, G., Gneezy, U., and Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior and Organization*, 189:234–256.

Danz, D., Vesterlund, L., and Wilson, A. J. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review*.

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015). Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism. *American Economic Review*, 105(11):3416–42.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Gino, F., Norton, M. I., and Weber, R. A. (2016). Motivated bayesians: Feeling moral while acting egoistically. *Journal of Economic Perspectives*, 30(3):189–212.

Gneezy, U. and Rustichini, A. (2000). Pay Enough or Don't Pay at All. *The Quarterly Journal of Economics*, 115(3):791–810.

Hossain, T. and Okui, R. (2013). The binarized scoring rule. *Review of Economic Studies*, 80(3):984–1001.

Insko, C. A. and Schopler, J. (1987). Categorization, competition, and collectivity.

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica*, 77(2):603–606.

Kugler, T., Bornstein, G., Kocher, M. G., and Sutter, M. (2007). Trust between individuals and groups: Groups are less trusting than individuals but just as trustworthy. *Journal of Economic Psychology*, 28(6):646–657.

Schlag, K. H., Tremewan, J., and van der Weele, J. J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, 18(3):457–490.

Schotter, A. and Trevino, I. (2014). Belief Elicitation in the Laboratory. *Annual Review of Economics*, 6(1):103–128.

Trautmann, S. T. and van de Kuilen, G. (2015). Belief elicitation: A horse race among truth serums. *The Economic Journal*, 125(589):2116–2135.

Chapter 1: Belief elicitation under competing motivations: Does it matter how you ask?*

Lata Gangadharan [†], Philip J. Grossman [‡], Nina Xue [§]

Abstract

Beliefs are increasingly recognised as an important driver of behaviour, but measuring beliefs is not straightforward. We design a giving experiment to compare beliefs (about others) using different elicitation mechanisms when self-serving motives may compete with accuracy incentives. Consistent with a simple theoretical framework, we find evidence of a self-serving bias for non-donors when beliefs are not incentivised, while donors' beliefs are more accurate, irrespective of the elicitation mechanism. Offering a simple incentive does not reduce non-donors' underestimation of actual giving, however, a variation of the Becker-DeGroot-Marschak (BDM) procedure does appear to mitigate the negative bias in beliefs by both structuring the belief question as a question about payment and increasing the salience of monetary incentives. Our results also show that biases in beliefs do not vary with the timing of belief elicitation.

JEL Classification: C9, D9, H4

Keywords: belief elicitation mechanisms, self-serving motive, donations, experiment

[*] This work was supported by the Australian Research Council (DP170101167). We are grateful to David Birke, Raphael Epperson, Ben Grodeck, Boon Han Koh, Jonathan Levy, Birendra Rai, Marta Serra-Garcia, Egon Tripodi, Erte Xiao, and participants at the Economic Science Association 2020 Global Around-the-Clock Virtual Meeting, the Southern Economic Association Annual Meeting (2020), the BEST conference on Human Behaviour & Decision Making (2021), the Social Image and Moral Behavior Workshop (2021), the Monash Mini-Workshop on Motivated Beliefs, the WZB Economics of Philanthropy Workshop, and the Innsbruck Winter Summit for valuable feedback. This project received approval from the Monash University Human Research Ethics Committee (project number 23704).

[†] Monash University, Victoria, Australia, Lata.Gangadharan@monash.edu

[‡] Monash University, Victoria, Australia, Philip.Grossman@monash.edu

[§] Monash University, Victoria, Australia, Nina.Xue@monash.edu

1 Introduction

Researchers are increasingly turning to beliefs to shed light on behavioural drivers. Two individuals with the same preferences may make vastly different decisions if they hold different beliefs. For example, an individual who believes a beggar is not in fact homeless but a “scam artist” might act differently to someone who does not share this scepticism, despite being equally altruistic. While useful in explaining behaviour, “true” beliefs, like preferences, cannot be directly observed. Previous work has evaluated various belief elicitation mechanisms based on the accuracy of beliefs relative to a Bayesian benchmark.[1] In a survey of belief elicitation mechanisms, Charness et al. (2021) conjecture that simple incentivised methods may outperform both non-incentivised introspection and more complex incentivised methods. The authors emphasise that there is little research directly comparing different elicitation mechanisms. Even less is known about how beliefs respond to different elicitation methods when beliefs may be motivated or biased by considerations other than accuracy.

The goal of this paper is to compare introspection with both a simple incentivised method and a more sophisticated incentivised method when self-serving motives may compete with incentives for belief accuracy. We present a simple theoretical framework of beliefs motivated by self-serving concerns. Agents face a trade-off between a desire to maximise monetary payoffs and minimise psychological costs (by holding accurate beliefs), against a desire to maximise self-image utility, or utility derived from norm compliance (by holding negatively biased beliefs about the generosity of others). We first investigate whether beliefs about others’ choices are biased when no incentive is offered. We then examine whether beliefs vary with the elicitation mechanism, via the relative salience of monetary and self-image utility.

To this end, we design a giving experiment in which participants can make a donation with a low probability of being implemented. We then elicit beliefs about the proportion of previous participants who chose to give, using one of three commonly used elicitation methods: non-incentivised (introspection), incentivised, and the incentivised Karni (2009) mechanism, a more sophisticated method that we present as a multiple price list.[2] We assume that individuals who are unbiased in the processing of information have rational expectations and that their beliefs about the proportion of donors will not systematically deviate from the true proportion (Di Tella et al., 2015). Any systematic variation from the empirical benchmark (i.e., the actual proportion of donors) at the aggregate level would indicate a bias in beliefs. We conjecture that a negative deviation from

[1] Trautmann and van de Kuilen (2015) present a “horse race” of various incentivised and non-incentivised mechanisms and find similar accuracy levels in beliefs, but the incentivised mechanisms are better predictors of actual behaviour. See also Schotter and Trevino (2014) and Schlag et al. (2015) for reviews of elicitation mechanisms.

[2] The method uses a direct revelation mechanism to elicit subjective probabilities, first introduced by Ducharme and Donnell (1973) as a variant of the Becker-DeGroot-Marschak mechanism (Becker et al., 1964) and later formalised theoretically by Karni (2009). The method is also known as the “bets mode”, “probability matching”, “reservation probabilities”, and “stochastic Becker-DeGroot-Marschak method”.

the benchmark is likely motivated by self-image concerns. Adopting the design from Gangadharan et al. (2023), participants whose donations were not initially implemented are offered a second chance, and can pay to increase the probability that the charity receives the donation. Altruistic motives are therefore increasing in the total amount that subjects are willing to pay to ensure that the donation is received.

In a charitable giving context, one advantage of using experimental methods is that data can be collected from both donors and non-donors, whereas observational data is typically not available for non-donors. Further, though surveys may provide some insight into individual beliefs, an experimental approach allows us to systematically compare non-incentivised beliefs against beliefs elicited using two popular incentive-compatible mechanisms.

First, we find that in the absence of incentives, individuals who choose not to give, systematically believe that others are also not generous, while donors' beliefs are substantially more accurate and do not deviate significantly from the empirical benchmark. We show that this belief gap between donors and non-donors is not explained by a pure consensus effect or by individual differences in optimism. Using data from three additional treatments, we further show that these belief biases are robust to the timing of belief elicitation, suggesting the existence of "non-giving types" who are not only consistent in choosing not to give (de Oliveira et al., 2011) but are also consistent in believing that others would not do so.

Our second result is that the belief biases in non-donors persist even after introducing a simple incentive. Under the more sophisticated Karni mechanism, however, non-donors' beliefs are substantially more accurate and approach the empirical benchmark. These findings are consistent with our theoretical framework which predicts that among the incentivised methods, monetary (self-image) utility is relatively more (less) salient under the Karni mechanism, thus it matters *how* you ask. In an additional treatment, we find that both the ability of the Karni mechanism to frame the belief question as a question about payment, and the greater complexity of the mechanism, play a role in mitigating belief biases. We also show that differences between the incentivised methods cannot be fully explained by the exclusion of inconsistent switchers or by cognitive uncertainty.

Our research makes several contributions. Our results highlight the need to choose elicitation mechanisms carefully, as different methods can trigger different motivations and as a consequence, produce different belief responses.[3] Simply offering a payment for beliefs is not sufficient to attenuate the negative bias in beliefs, but more sophisticated incentive mechanisms such as the Karni mechanism could wash out other motivations as monetary concerns are made more salient. Within non-giving types, we identify biased beliefs about others that persist irrespective of the timing of belief elicitation. This is economically relevant for organisations and policymakers and suggests an alternative avenue for encouraging prosocial behaviour – by focusing on debiasing inaccurate

[3] Danz et al. (2022) find that beliefs are less accurate under full information about the payment mechanism and highlight the role of incentives in distorting reported beliefs.

beliefs, rather than by attempting to change underlying preferences. Previous studies have found that providing accurate information about the behaviour of others can be effective at changing behaviour (e.g., Shang and Croson, 2009; Dimant and Gesche, 2021; Bicchieri et al., 2021). Our findings offer a reason for their effectiveness, i.e., by making it more costly for non-giving types to both choose selfishly and maintain a positive self-view.

The following section relates our paper to the existing literature. Section 3 presents the experimental design. In Section 4, we present a simple theoretical framework of beliefs motivated by self-image and our main hypotheses. The main results are reported in Section 5. In Section 6, we introduce five additional treatments as robustness checks and explore the plausibility of alternative channels to explain our results. Finally, Section 7 concludes.

2 Related Literature

Our research connects to two main strands of the literature. First, we build on a recent body of work on the measurement of beliefs. Second, our paper is related to a growing literature on motivated beliefs.

Belief elicitation

One obvious way to elicit beliefs is to simply ask, without any incentives, also known as introspection.[4] Though straightforward, a drawback of this mechanism is that individuals may not think carefully enough about their answer, may receive an expressive value from reporting a particular view (e.g., Bullock et al., 2013), or may fall prey to a hypothetical bias (e.g., List and Gallet, 2001). Experimentalists have tried to address these concerns by making belief revelation incentive-compatible, compelling agents to make a trade-off between financial and non-financial motivations. There is, however, ample experimental evidence that individuals are willing to forgo monetary gains to satisfy other preferences, and even very high stakes may not be sufficient to eliminate cognitive biases (e.g., Enke et al., 2021). Coutts (2019) offers evidence that higher payments for accuracy can increase belief biases in the presence of anticipatory utility.[5] On the other hand, Zimmermann (2020) finds that large incentives can improve the ability to recall negative feedback.

Previous work has examined the interaction between risk preferences and elicitation mechanisms, leading to the popularity of the Karni mechanism and the binarized scoring rule (BSR), due to their invariance to heterogeneous risk preferences.[6] Danz et al. (2022) show that provid-

[4] Baillon et al. (2022) find no difference between hypothetical and incentivised responses in the absence of defaults, but that incentives can reduce the default bias.

[5] Coutts (2019) compares beliefs elicited using the Karni mechanism against beliefs elicited using a simple incentive. However, the two beliefs also differ in whether incentives exist for belief distortion, making it difficult to directly compare the methods.

[6] We chose the Karni mechanism as a comparison against a simple incentivised mechanism because of this property and its increasing popularity in the literature. The interaction between BSR and risk preferences is reported in Hossain

ing detailed information about the BSR reduces both belief accuracy and the explanatory power of beliefs for behaviour as beliefs no longer explain differences in behaviour in the Niederle and Vesterlund (2007) task. Burfurd and Wilkening (2021) explore the interaction between elicitation mechanisms and cognitive ability and find that, compared to no incentive, the Karni mechanism results in larger differences in belief accuracy between subjects with low and high cognitive ability. To the best of our knowledge, discussions around elicitation mechanisms focus on the accuracy of belief updating (against a Bayesian benchmark), and have so far neglected the interaction between different elicitation methods and beliefs that are potentially biased by self-serving concerns, which is the key objective of this paper.

Motivated beliefs

Motivated beliefs result from a set of biased cognitive processes related to the gathering, processing, and recall of information (e.g., Kunda, 1990). In economics, motivated reasoning implies a preference over particular beliefs (e.g., Bénabou and Tirole, 2016), while psychologists reason that there are multiple, and often conflicting, motivations that are competing for one's attention (e.g., Epley and Gilovich, 2016). Gino et al. (2016) posit that individuals have a preference over a positive self-image, i.e., a preference for *feeling* moral without necessarily incurring the costs associated with *being* moral. These "Motivated Bayesians" require some degree of mental flexibility in order to hold and maintain motivated beliefs. Chen and Heese (2021) find support for this in their experiment, as individuals with above-average cognitive ability are more likely to acquire information in a self-serving manner.[7]

Motivated beliefs often go hand-in-hand with excuse-driven selfishness.[8] While there is increasing evidence that belief biases are stronger for individuals who make more selfish choices (e.g., Molnár and Heintz, 2016; Serra-Garcia and Szech, 2021; Andreoni and Sanchez, 2020), belief distortions and subsequent excuse-driven selfishness do not always occur (e.g., Van der Weele et al., 2014; Bartling and Özdemir, 2022; Valero, 2021). Iriberri and Rey-Biel (2013) find a positive correlation between giving and beliefs about the generosity of others that appears to be strongest for selfish types. Di Tella et al. (2015) present a variant of the dictator game, in which receivers can accept a side payment to reduce the total endowment. Dictators who are able to take more for themselves are more likely to believe the receiver was selfish and this self-serving bias persists even in the presence of a large monetary incentive for correct beliefs. Bicchieri et al. (2020) find evidence of distorted beliefs about descriptive norms and subsequently observe higher rates of selfish

and Okui (2013) and Erkal et al. (2020).

[7] Such self-serving biases can have an instrumental value, for example, overconfidence can be useful in influencing others in social interactions (e.g., Schwardmann et al., 2022; Solda et al., 2020)

[8] Excuse-driven selfishness is prevalent in a variety of domains including situations with moral "wiggle room" (Dana et al., 2007), situations in which strategic ignorance or inattention is possible (e.g., Exley and Petrie, 2018; Grossman and van der Weele, 2017) and in the presence of uncertainty (e.g., Exley, 2016; Haisley and Weber, 2010).

behavior. Similarly in Ging-Jehli et al. (2020), subjects who take more from another participant are more likely to believe the other participant was selfish, however, the authors find that overall, beliefs are not significantly different between second parties and third party observers. Given the mixed evidence, it is important to better understand *when* beliefs are more likely to be biased.[9] We investigate whether the identification of self-serving beliefs depends on the elicitation mechanism, by comparing introspection to a simple incentive and a more complex method.

3 Experimental design

We design a between-subjects experiment with three treatments, varying the mechanism used to elicit beliefs. The experiment consists of three stages with subjects receiving the instructions for each stage only after completing the preceding stage (see Appendix E for the instructions). In Stage 1, participants can donate to a charity with a low probability that the donation is implemented. We introduce a probabilistic donation in order to identify altruistic concerns (Gangadharan et al., 2023), based on willingness to pay to increase donation probability in Stage 3 (which comes as a surprise). In Stage 2, we elicit beliefs about the proportion of donors using one of three elicitation mechanisms. These beliefs can also be interpreted as empirical expectations about social norms (Bicchieri, 2005). We chose these beliefs based on previous work showing the importance of empirical expectations in predicting prosocial behaviour (e.g., Bicchieri and Xiao, 2009; Bicchieri et al., 2020, 2021; Danilov et al., 2021). In the anonymised context of our experiment, norm violations are not observable by others. Therefore, disutility from not complying with the norm would most likely be related to self-image rather than social image (with the norm being "internalised").

Stage 1: Donation decision

In Stage 1, participants complete a real-effort task consisting of questions from Raven's Progressive Matrices (Raven and Court, 1938), and receive a fixed endowment plus a piece-rate for every correct answer (to encourage effort). This provides a proxy for cognitive ability, which previous work suggests could be correlated with motivated reasoning (e.g., Gino et al., 2016; Chen and Heese, 2021). Participants choose a charity that they believe is most worthy from a list provided and then have the option of donating a small portion (x) of their endowment (Y) to this charity, with a probability $p = 0.10$ that the donation is implemented (i.e., from 10 cards displayed, 9 red and 1 green, the green card is drawn), in which case the experimenter matches the amount and $2x$ is donated. If a red card is drawn, the donation is not implemented. Participants are informed of the draw immediately after making their donation decision. In order to increase the donation rate and to have a sufficient sample of donors for Stage 3, we chose a small donation amount and

[9] Drobner (2022) offers a step in this direction, showing that beliefs are more likely to be biased when individuals are not expecting to receive feedback.

a low probability of implementation, thereby keeping the expected price of giving low (Andreoni and Miller, 2002).

Stage 2: Belief elicitation

In Stage 2, we ask for beliefs regarding others' donations using one of three commonly used mechanisms: non-incentivised (*NonInc*), incentivised (*Inc*), or Karni (*Inc-Karni*). In *NonInc* and *Inc*, participants are informed that a previous group of 10 participants faced the same donation decision that they had just encountered. Participants are asked to guess how many of the previous participants they think chose to give. In *Inc*, participants receive an additional amount if they correctly guess the actual number of donors. As we explain below, we chose this amount such that it is equal to the donation amount (x).

In *Inc-Karni*, the probability that the participant receives the additional payment is increasing in the accuracy of beliefs. We present the Karni mechanism as a multiple price list (see Table 1), following Trautmann and van de Kuilen (2015). One major advantage of this format is that it allows the belief question to be structured as a question about payment (Andreoni and Sanchez, 2020), which cannot be achieved with introspection, a simple binary incentive or a direct lottery-based presentation of the Karni method.[10] Previous work has shown that the choice menu presentation of the Karni method results in fewer boundary reports than the standard BDM format (Holt and Smith, 2016). Freeman and Mayraz (2019) also suggest that choice lists may help scaffold decision making and lead to more informed decisions. Participants choose between two options in 11 scenarios, with one scenario selected at random for payment. Option A corresponds to the amount given by a previous participant (i.e., x if they chose to donate, and zero otherwise), to be paid by the experimenter. This is the same across all 11 Scenarios. Option B is an outside gamble in which participants receive x with probability ranging from 0% to 100% in steps of 10%, and zero with probability ranging from 100% to 0% in steps of 10%.[11] We can deduce subjective beliefs by observing when a participant switches from Option A to Option B.[12] For example, a subject who believes there is a 65% chance that a previous subject chose to donate would maximise their expected payoff by switching from Option A to Option B at Scenario 8. If they switched earlier, e.g., at Scenario 7, then according to their belief, Option A gives them a 65% chance of receiving x, while Option B only gives them a 60% chance. In other words, the subject foregoes an additional 5% chance of receiving x.

We conducted additional treatments to explore whether the beliefs we elicit are robust to the

[10] We discuss the significance of this in relation to the theoretical framework in Section 4.

[11] To keep Option A and B consistent, the belief payment is the same as the amount a subject would choose to donate (x).

[12] We use wording from Exley's (2016) normalization price list by informing subjects "Most people begin by preferring Option A and then switch to Option B." We do not enforce a single switching point in order to identify subjects who may be confused or have other preferences.

timing of belief elicitation (either before or after the donation ask). We discuss these treatments in more detail in Section 6.2.

Table 1: The Karni mechanism presented as a multiple price list

Scenario	**Option A**: Amount given by previous subject (0 or x)	**Option B**: lottery with different chances of receiving 0 and x
1	Amount given by previous subject	(0 with 100%), (x with 0%)
2	Amount given by previous subject	(0 with 90%), (x with 10%)
3	Amount given by previous subject	(0 with 80%), (x with 20%)
4	Amount given by previous subject	(0 with 70%), (x with 30%)
5	Amount given by previous subject	(0 with 60%), (x with 40%)
6	Amount given by previous subject	(0 with 50%), (x with 50%)
7	Amount given by previous subject	(0 with 40%), (x with 60%)
8	Amount given by previous subject	(0 with 30%), (x with 70%)
9	Amount given by previous subject	(0 with 20%), (x with 80%)
10	Amount given by previous subject	(0 with 10%), (x with 90%)
11	Amount given by previous subject	(0 with 0%), (x with 100%)

Stage 3: Second donation decision and survey

For participants whose donations were not implemented in Stage 1, Stage 3 offers a second chance. Participants can spend an additional amount (a) to increase the implementation probability (i.e., increase (reduce) the number of green (red) cards and draw another card).

As an alternative to a binary classification of giving, we use a more continuous measure to gauge the strength of altruistic concerns, see Gangadharan et al. (2023) on the experimental method and validation of the method using an existing survey measure (Carpenter, 2021). This procedure allows us to further classify donors based on the relative strength of their altruistic motives, i.e., how much they spend to increase the probability.[13] Following Stage 3, participants completed a survey with socio-demographic questions on gender, age, education, religiosity, political ideology and income. Subjects were only informed about their final payoffs upon completing the survey.

3.1 Procedures

The experiment was programmed in oTree (Chen et al., 2016) and was conducted on Amazon Mechanical Turk (MTurk) between May-October 2020 with 350 participants across *NonInc* ($N =$

[13] A key distinction between altruistic and warm-glow giving is that warm-glow utility is derived as soon as a giving decision is made, whereas altruistic utility depends on the outcome for the recipient (e.g., Andreoni, 1989; Null, 2011; Ottoni-Wilhelm et al., 2017; Gangadharan et al., 2018; Tonin and Vlassopoulos, 2013; Andreoni and Serra-Garcia, 2021).

100), *Inc* ($N = 102$) and *Inc-Karni* ($N = 148$).[14] Previous studies have shown that the behaviour of participants is comparable between the lab and MTurk and that the results of online experiments can be generalised to both the lab and field (e.g., Horton et al., 2011; Snowberg and Yariv, 2021).

Participants received an endowment of $Y =$ US\$2.50, and a piece-rate of \$0.10 for every correct answer in Stage 1. The initial donation cost participants $x = \$0.40$, and for every a spent, the implementation probability increased by $a \cdot p/3$ in Stage 3 (i.e., every $a = \$0.03$ corresponds to an increase in probability of 10%). In *Inc* and *Inc-Karni*, subjects could receive an additional $x = \$0.40$ based on belief accuracy in Stage 2. Figure 1 summarises the experimental procedure. Decisions were anonymous and participants earned an average of US\$2.74, for a median completion time of 13 minutes, equivalent to approx. US\$12.65 per hour which is well above the average hourly wage on MTurk (e.g., Hara et al., 2018).[15] Consistent with previous studies using a multiple price list format (e.g., Möbius et al., 2022; Dave et al., 2010; Bandyopadhyay et al., 2021), we excluded 22% of participants from *Inc-Karni* ($N = 33$, with $N = 115$ remaining) due to multiple switching or switching in the opposite direction in Stage 2, making it difficult to determine their belief. Section 6.1 discusses this further with robustness checks.

4 Beliefs motivated by self-serving concerns

In the spirit of Bodner and Prelec (2003) and Bénabou and Tirole (2006), we outline a simple theoretical framework that we draw upon to develop the testable hypotheses relating to the beliefs of participants. In Stage 1, the agent makes their donation decision, $X \in \{0, x\}$. The true proportion of donors in our sample (our empirical benchmark) is given by $\lambda \in [0, 1]$ while the individual's belief about this proportion is denoted by $\hat{\lambda} \in [0, 1]$. Note that these beliefs are about how *others* behave. In Stage 2, the agent can earn an additional payment (x), based on their reported belief. We assume that if agents are rational in the processing of information and beliefs are unbiased, then expectations about the proportion of donors will not deviate from the true proportion, $\hat{\lambda} = \lambda$.[16]

The incentive for belief accuracy is represented by $m(\hat{\lambda}, \lambda)$, in which the probability of receiving the belief payment is decreasing in the difference between $\hat{\lambda}$ and λ and is concave. Comparing across the incentivised treatments, the cost of reporting an inaccurate belief is higher in *Inc* than in *Inc-Karni*.[17] To see this intuition, at an extreme, when $\lambda = 1$ and $\hat{\lambda} < \lambda$, agents in *Inc* forgo the

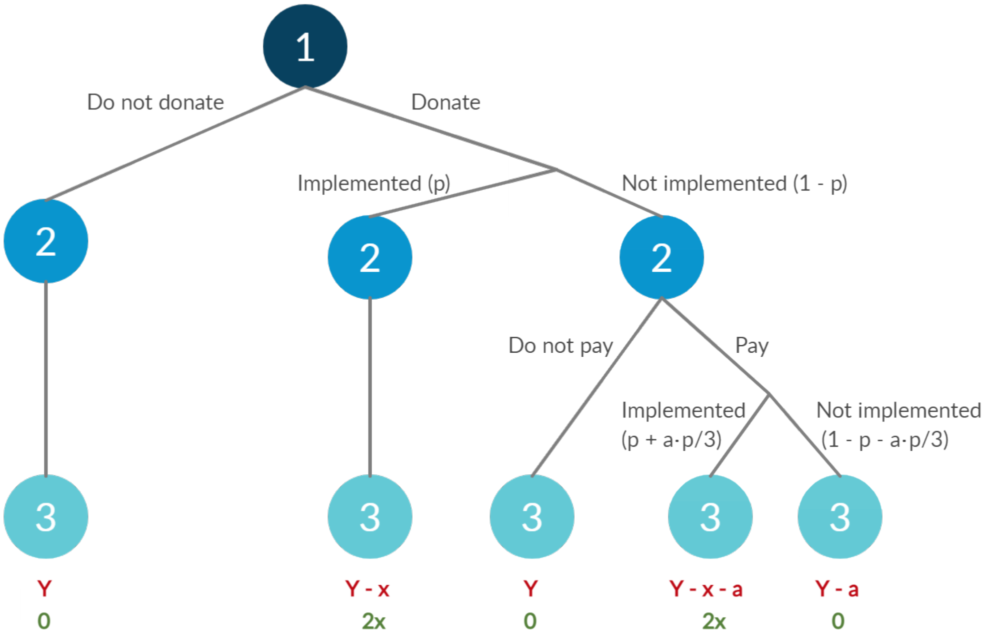
[14] Based on pilot data, this allows us to detect an effect size of 0.96 standard deviations in beliefs, with 80% power and a Type I error rate of 95%.

[15] To improve the quality of data collected, we restricted participation to individuals located in the United States with a high approval rate in their previously completed Human Intelligence Tasks (HITs) and included comprehension questions.

[16] See Di Tella et al. (2015) for a similar assumption.

[17] For example, suppose $\lambda = 0.5$ and $\hat{\lambda} = 0.2$. In *Inc*, reporting $\hat{\lambda} = 0.5$ would result in a 25% chance of receiving x (based on a binomial distribution, as payment is based on a random sample of 10 previous participants making a binary donation decision) while reporting $\hat{\lambda} = 0.2$ only results in a 4% chance. In *Inc-Karni*, reporting $\hat{\lambda} = 0.5$ results in a 64% probability, while reporting $\hat{\lambda} = 0.2$ results in a slightly lower probability of 61%.

Figure 1: Experimental procedure



Notes: Stage 1: Real-effort task and donation decision with probability $p = 0.10$ of implementation. Stage 2: Belief elicitation. Stage 3: Second donation decision (for a subset) with probability $(p + a \cdot p/3)$ of implementation. The decision-maker's payoff is presented in the top row while the charity's payoff is denoted in the bottom row. $Y = \$2.50$ denotes the participant's endowment, $x = \$0.40$ denotes the donation amount, and a denotes the amount paid to increase the probability (every $a = \$0.03$ corresponds to a 10% increase).

belief payment with certainty, while agents in *Inc-Karni* only forgo some probability of earning this payment.

In addition to a potential financial cost of holding biased beliefs, we assume that there is also a psychological cost, $c(\hat{\lambda}, \lambda)$, that is increasing in the difference between $\hat{\lambda}$ and λ and is strictly convex. This follows Kunda (1990), who argues that beliefs are motivated to the extent that an individual can convince a third party of their beliefs. The larger the belief bias, the larger the psychological cost and the more difficult it is to convince a reasonable third party that the belief is an accurate one.[18] In *NonInc*, despite there being no financial penalty for inaccurate beliefs, the agent is nonetheless constrained by these psychological costs.

To represent self-serving concerns, we assume that agents have uncertainty about whether they are a prosocial or selfish type, and derive self-image (or ego) utility, $E[\theta]$, from attaching a probability of $\theta \in [0, 1]$ to being the prosocial type.[19] Given that the donation decision is a binary

[18] Another interpretation would be the additional cognitive effort required to selectively recall and process information.

[19] A complementary interpretation is that agents derive utility from norm compliance (by choosing $X = x$), or

one, we conjecture that donors, having given the maximum amount possible, derive sufficient self-image utility through their donation ($X = x$) and thus have no need to bias their beliefs about others, as the probability of them being a prosocial type is already sufficiently high based on their action. Beliefs, therefore, do not enter into self-image utility for donors.[20] Non-donors, on the other hand, are unable to derive the same self-image utlity through their actions as we assume that $E[\theta|X = x] > E[\theta|X = 0]$, i.e., self-image utility is higher for donors than non-donors.[21] Thus, non-donors can only protect their self-image by believing that most others in the same position also would not donate. This downward distortion of beliefs (about others) renders the agent's own decision not to donate less informative about their type. For non-donors, self-image utility is therefore decreasing in their belief about the generosity of others. An individual's belief decision is modelled by:

$$\max_{\hat{\lambda}\in[0,1]} \beta \cdot m(\hat{\lambda}, \lambda) - c(\hat{\lambda}, \lambda) + \mu \cdot (\mathbf{1}_D \cdot E[\theta|X = x] + (1 - \mathbf{1}_D) \cdot E[\theta|X = 0, \hat{\lambda}]) \tag{1}$$

where $\mathbf{1}_D$ takes a value of 1 for donors, and 0 otherwise. The weight that individuals place on money is represented by β while μ is the weight assigned to self-image (i.e., how much agents care about being the prosocial type). For non-donors, our stylised model captures the tension between a desire to maximise financial payoffs ($m(\hat{\lambda}, \lambda)$) and minimise psychological costs ($c(\hat{\lambda}, \lambda)$), against a desire to maximise self-image utility ($E[\theta|X = 0, \hat{\lambda}]$). Taking the first order condition with respect to $\hat{\lambda}$ for the interior solution yields:

$$\begin{cases} \beta \cdot m'(\hat{\lambda}^*, \lambda) - c'(\hat{\lambda}^*, \lambda) + \mu \cdot E'[\theta|X, \hat{\lambda}^*] = 0, & \text{if } \mathbf{1}_D = 0 \\ \beta \cdot m'(\hat{\lambda}^*, \lambda) - c'(\hat{\lambda}^*, \lambda) = 0, & \text{if } \mathbf{1}_D = 1 \end{cases} \tag{2}$$

In Hypothesis 1, we first examine beliefs in the absence of an incentive (the first component in (2) disappears). Among non-donors in *NonInc*, assuming that the psychological costs are small relative to the potential gains in self-image utility, beliefs will be biased in a downward direction. It is straightforward to see that for donors, the optimal belief is simply one that minimises the psychological costs, i.e., $\hat{\lambda}^* = \lambda$.

Hypothesis 1 (*Self-serving bias*) In *NonInc*, non-donors' beliefs are lower than the true proportion of donors, while donors' beliefs are not significantly different from this empirical benchmark.

derive disutility from not complying with the social norm (by choosing $X = 0$). See Bicchieri (2005) for a norm-based utility framework.

[20] Our results would hold even if we relax this assumption as we assume that the marginal benefits for self image are negligible for donors compared to the marginal costs of belief distortion (both psychologically and for their payoff).

[21] An alternative interpretation of the donation decision is that it is a proxy for whether the agent is a giving or non-giving type (de Oliveira et al., 2011). In particular, not willing to donate in our study when it is relatively cheap to do so is a strong indicator that an individual is a non-giving type.

$$\begin{cases} \hat{\lambda}^{NonInc} < \lambda, & \text{if } \mathbf{1}_D = 0 \\ \hat{\lambda}^{NonInc} = \lambda, & \text{if } \mathbf{1}_D = 1 \end{cases} \tag{3}$$

In Hypothesis 2a and 2b, we compare beliefs across the different elicitation methods. We expect the presence of a monetary incentive to reduce biases in beliefs. Comparing across the incentivised mechanisms, the standard economic prediction is that beliefs will be less biased in *Inc* than *Inc-Karni* because the relative cost of reporting an inaccurate belief is higher in the former. For donors, we do not expect beliefs to vary across the three methods.

Hypothesis 2a (*Monetary costs*) Non-donors' beliefs are lowest in *NonInc*, followed by *Inc-Karni*, and finally *Inc*. Donors' beliefs do not depend on the elicitation mechanism.

$$\begin{cases} \hat{\lambda}^{NonInc} < \hat{\lambda}^{Inc-Karni} < \hat{\lambda}^{Inc}, & \text{if } \mathbf{1}_D = 0 \\ \hat{\lambda}^{NonInc} = \hat{\lambda}^{Inc-Karni} = \hat{\lambda}^{Inc}, & \text{if } \mathbf{1}_D = 1 \end{cases} \tag{4}$$

An alternative behavioural hypothesis is that the way in which beliefs are elicited affects the relative weights placed on payoff and image utility. Assuming that $\hat{\lambda} < \lambda$ for non-donors, ceteris paribus, an increase in β will increase $\hat{\lambda}^*$ while an increase in μ will decrease $\hat{\lambda}^*$. In other words, increasing the salience of monetary incentives will place upward pressure on beliefs towards the benchmark, while increasing the salience of self-image will put downward pressure on beliefs away from the benchmark. The importance of salience in helping agents allocate limited cognitive resources has been studied in both psychology (e.g., Taylor and Thompson, 1982) and economics (e.g., Bordalo et al., 2012, 2013; Gabaix, 2019). In contrast to standard economic theory, decisions may differ depending on the framing of the decision problem, as some aspects are made more psychologically salient than others. We conjecture that the ability of the Karni mechanism to frame the belief question as a question about payment, coupled with its greater complexity, will increase the relative salience of monetary utility (β) and decrease the relative salience of self-image utility (μ), resulting in smaller belief biases in *Inc-Karni*. Comparing across *NonInc* and *Inc* gives us the effect of an increase in the *level* of the incentive (from \$0.00 to \$0.40). Note that we hold the *magnitude* of the incentive constant (at \$0.40) in *Inc* and *Inc-Karni*, and only vary the psychological *salience* of the monetary incentive.

Hypothesis 2b (*Salience*) Non-donors' beliefs are lowest in *NonInc*, followed by *Inc*, and finally *Inc-Karni*. Donors' beliefs do not depend on the elicitation mechanism.

$$\begin{cases} \hat{\lambda}^{NonInc} < \hat{\lambda}^{Inc} < \hat{\lambda}^{Inc-Karni}, & \text{if } \mathbf{1}_D = 0 \\ \hat{\lambda}^{NonInc} = \hat{\lambda}^{Inc} = \hat{\lambda}^{Inc-Karni}, & \text{if } \mathbf{1}_D = 1 \end{cases} \tag{5}$$

Using the experimental measure introduced by Gangadharan et al. (2023) to identify the strength of altruistic motives, we check whether our results relating to the hypotheses above are robust to a more continuous measure of prosocial preferences.

5 Results

On average, participants reported a belief that 4.99 (std. dev. = 2.74) out of the 10 participants from a previous session chose to donate. When given the option to donate, 57% chose to give (our empirical benchmark) and donation rates did not differ significantly across treatments at the 5% level (see Appendix A).[22] Of the donors who did not have their initial donation implemented, 60% paid to increase the probability of implementation in Stage 3, on average increasing the implementation probability to 40%. All results reported below hold when we exclude donors whose initial donations were implemented (see Appendix B). We next report our findings relating to each of our hypotheses.

5.1 Non-donor and donor beliefs in *NonInc*

We first examine beliefs in the absence of an incentive. On average, non-donors reported a belief that 2.94 out of 10 previous participants donated, which is 47% lower than the true proportion (one-tailed Wilcoxon signed-rank test, $p < 0.01$). Donors on the other hand, reported an average belief of 5.90, which is not significantly different from the empirical benchmark (one-tailed Wilcoxon signed-rank test, $p > 0.10$). Using a one-tailed Mann-Whitney test (unless otherwise specified, we use one-tailed Mann-Whitney tests to compare mean beliefs), we find that the belief gap between donors and non-donors is significantly greater than zero ($p < 0.01$). This result is robust to the inclusion of demographic controls and accounting for multiple hypothesis testing using the Bonferroni correction in the OLS regression analysis in Table 2 ($p < 0.01$, column 2), and offers support for a self-serving bias in beliefs among non-donors, consistent with Hypothesis 1.

Result 1: In *NonInc*, non-donors underestimate the true proportion of donors, while donors' beliefs do not deviate significantly from the empirical benchmark.

[22] We use the average donation rate for all participants as the empirical benchmark, rather than the donation behaviour of the small sample of 10 previous participants.

Table 2: Beliefs in *NonInc*

	(1)	(2)
Non-Donor	-2.96^{***}	-2.83^{***}
	(0.48)	(0.50)
Raven's score		0.13
		(0.17)
Constant	5.90^{***}	5.62^{***}
	(0.34)	(1.35)
Controls	*No*	*Yes*
R^2	0.28	0.46
Adj. R^2	0.27	0.33
Num. obs.	100	100

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

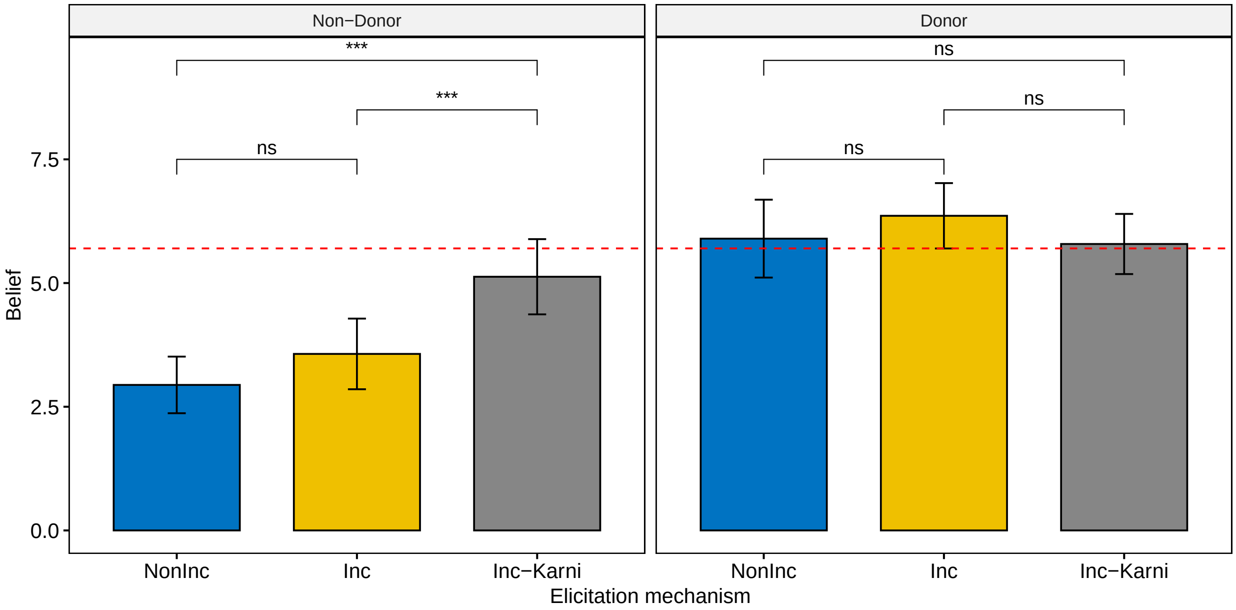
Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the subject's belief about the proportion of donors. Statistical significance accounts for multiple hypothesis testing (adjusted using the Bonferroni correction). The control variables are: gender, age, education, religiosity, political ideology and income.

5.2 Beliefs across elicitation mechanisms

Next, we investigate whether different belief elicitation mechanisms have differing effects on the belief response. Figure 2 presents a comparison of mean beliefs across *NonInc*, *Inc* and *Inc-Karni*, for donors and non-donors. Qualitatively, non-donors' beliefs are lowest in *NonInc*, followed by *Inc*, and finally *Inc-Karni*, which is more consistent with Hypothesis 2b than 2a. Using a one-sided Jonckheere-Terpstra (JT) test, we find a significant ascending order for non-donors ($p < 0.01$). Surprisingly, we do not find a significant difference in beliefs between *NonInc* and *Inc* (2.94 vs. 3.57, $p > 0.10$). This suggests that simply offering an incentive for beliefs does not necessarily improve belief accuracy. However, we do find that non-donors' beliefs in *Inc-Karni* are higher than both *NonInc* (5.21 vs. 2.94, $p < 0.01$) and *Inc* (5.21 vs. 3.57, $p < 0.01$). This is in line with our conjecture that the combination of monetary incentives and salience is important in reducing belief biases in *Inc-Karni*. Donors' beliefs, on the other hand, do not differ significantly between *NonInc* and *Inc* (5.90 vs. 6.29, $p > 0.10$), nor do they differ between *Inc* and *Inc-Karni* (6.29 vs. 5.28, $p > 0.10$). We do not find a significant ascending order for donors (JT test, $p > 0.10$).

These results hold after controlling for demographic variables in the regression analysis (Table 3). Columns 2 and 4 show that consistent with the results reported above, non-donors' beliefs are higher in *Inc-Karni* than both *NonInc* and *Inc*, while donors' beliefs do not differ. Columns 5 and 6 pool data for all subjects and we find a significantly positive interaction between *Inc-Karni* and non-donors ($p < 0.01$), which all but cancels out the belief gap between donors and non-donors. While non-donors' scores in the cognitive ability test appear to be negatively correlated with beliefs and donors' scores seem to positively predict beliefs, these coefficients are not significantly different

Figure 2: Beliefs by elicitation mechanism for donors and non-donors



Notes: Mann-Whitney test, error bars represent standard errors. Dotted line represents the empirical benchmark. *** denotes $p < 0.01$; ** denotes $p < 0.05$; * denotes $p < 0.10$; ns denotes $p > 0.10$.

from zero ($p > 0.10$). Contrary to the results reported by Chen and Heese (2021), we do not find sufficient evidence that cognitive ability is negatively correlated with the beliefs of non-donors.

To examine whether our main results hold using a more continuous measure (as opposed to a binary measure based on a single donation choice), we use the experimental measure by Gangadharan et al. (2023) to identify the strength of altruistic motives. Among our sample, 43% chose not to donate in Stage 1, 20% made an initial donation in Stage 1 only, and 30% donated in Stage 1 and paid to increase the probability of the donation being implemented in Stage 3.[23] We therefore obtain a more fine-grained measure by examining the total amount a subject is willing to pay to increase the probability that the donation is implemented. For donors, we take the sum of the initial donation in Stage 1 and the amount paid in Stage 3. For non-donors, this variable takes a value of zero.

Table 4 shows that beliefs are significantly higher as the total amount paid increases ($p < 0.01$, column 2). Similar to Result 2, these biases are attenuated in *Inc-Karni*, as indicated by the negative coefficient of the interaction term ($p < 0.01$, column 2). This confirms our previous finding that while those with weaker altruistic concerns are better able to distort their beliefs under *NonInc* and *Inc*, these biases are substantially smaller in *Inc-Karni*. Thus, using an alternative procedure for measuring the strength of altruistic concerns, we find further evidence that less altruistic types

[23] For the remaining 7%, donations were implemented in Stage 1.

Table 3: Beliefs of donors and non-donors

	Non-Donors		Donors		Pooled	
	(1)	(2)	(3)	(4)	(5)	(6)
Inc	0.63	0.45	0.46	0.53	0.46	0.39
	(0.52)	(0.58)	(0.49)	(0.53)	(0.47)	(0.50)
Inc-Karni	2.19^{***}	2.08^{***}	-0.11	-0.07	-0.11	-0.14
	(0.46)	(0.50)	(0.50)	(0.55)	(0.48)	(0.52)
Raven's score		-0.13		0.08		-0.04
		(0.15)		(0.15)		(0.10)
Non-Donor					-2.96^{***}	-2.80^{***}
					(0.50)	(0.52)
Inc x Non-Donor					0.16	0.13
					(0.71)	(0.76)
Inc-Karni x Non-Donor					2.29^{***}	2.33^{***}
					(0.68)	(0.71)
Constant	2.94^{***}	2.74	5.90^{***}	4.19^{**}	5.90^{***}	4.98^{***}
	(0.33)	(1.23)	(0.37)	(1.37)	(0.35)	(0.95)
H_0: *Inc* = *Inc-Karni*	$p < 0.01$	$p < 0.01$	$p = 0.22$	$p = 0.24$	$p = 0.21$	$p = 0.27$
Controls	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
R^2	0.14	0.26	0.01	0.12	0.20	0.23
Adj. R^2	0.13	0.13	-0.00	-0.01	0.19	0.17
Num. obs.	143	143	170	170	313	313

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the subject's belief about the proportion of donors. The baseline Treatment is *NonInc*. Statistical significance accounts for multiple hypothesis testing (adjusted using the Bonferroni correction). The control variables are: gender, age, education, religiosity, political ideology and income.

are prone to belief biases but that this is mitigated in *Inc-Karni*.

Result 2: Non-donors' beliefs are lowest in *NonInc* and simply offering an incentive does not change the belief response. However, non-donors' beliefs are significantly higher in *Inc-Karni*. Donors' beliefs do not vary with the elicitation mechanism.

6 Alternative explanations and robustness checks

To examine the robustness of our findings and consider alternative explanations, we conduct further analysis and report results from five additional treatments with data from a total of 704 participants.

Table 4: Beliefs by the strength of altruistic motivations

	(1)	(2)
Altruism	7.16^{***}	6.94^{***}
	(1.09)	(1.14)
Inc	0.63	0.55
	(0.52)	(0.56)
Inc-Karni	2.29^{***}	2.31^{***}
	(0.47)	(0.50)
Altruism x *Inc*	-1.10	-1.24
	(1.57)	(1.68)
Altruism x *Inc-Karni*	-6.20^{***}	-6.39^{***}
	(1.48)	(1.55)
Raven's score		-0.02
		(0.11)
Constant	2.88^{***}	2.27^{*}
	(0.34)	(0.91)
H_0: *Inc* = *Inc-Karni*	$p < 0.01$	$p < 0.01$
Controls	*No*	*Yes*
R^2	0.22	0.25
Adj. R^2	0.21	0.18
Num. obs.	293	293

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

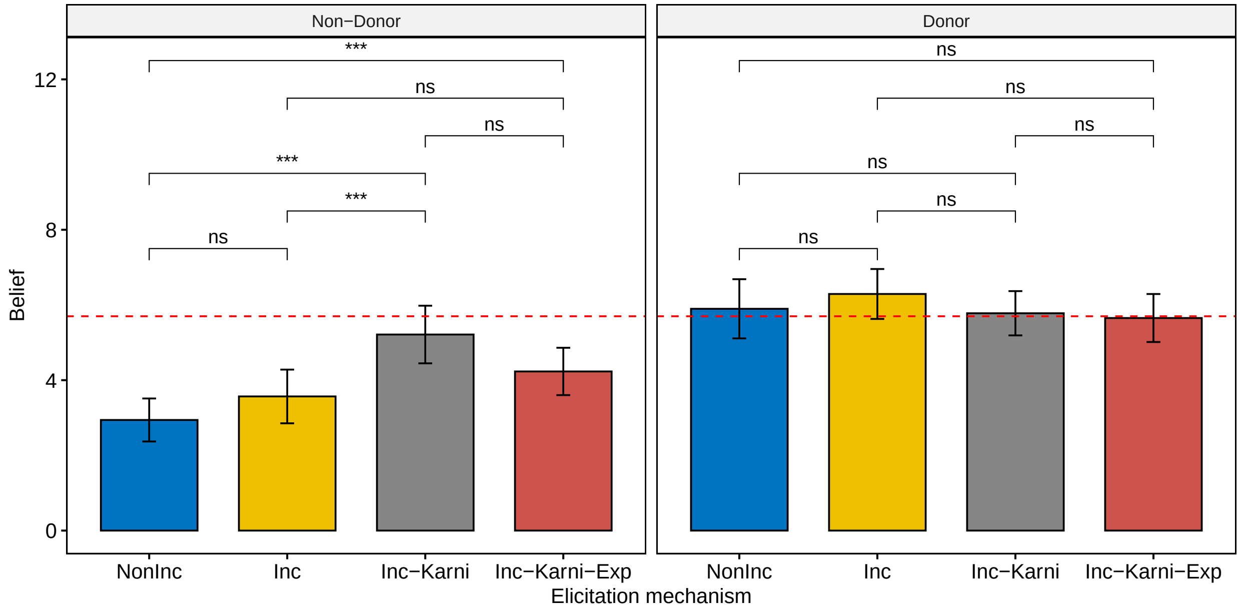
Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the subject's belief about average generosity. The baseline Treatment is *NonInc*. The strength of altruistic motivations is measured by the amount paid in Stage 1 ($0.40) and Stage 3 ($0.00 to $0.67). Donors whose initial donations were implemented are excluded. Statistical significance accounts for multiple hypothesis testing (adjusted using the Bonferroni correction). The control variables are: gender, age, education, religiosity, political ideology and income.

6.1 Why are beliefs different under Karni?

In this section we delve deeper into the mechanism driving differences between *Inc-Karni* and *Inc* and present results from an additional treatment (*Inc-Karni-Exp*) which highlight the importance of both the ability of the Karni mechanism to frame the belief question as a question about payment, and the mechanism itself in mitigating belief biases. We also assess potential explanations for why beliefs appear to be less biased under the Karni mechanism, supported by survey evidence from an additional treatment (*Inc-Karni-Survey*).

Salience under *Inc-Karni*: There are two main differences between *Inc* and *Inc-Karni*. First, the Karni mechanism, when presented in a multiple price list format, enables the belief question to be framed as a question about payment (Andreoni and Sanchez, 2020). Second, the mechanisms themselves differ in the way in which beliefs are incentivised. We conducted an additional treatment, *Inc-Karni-Exp* ($N = 128$), to disentangle these two explanations.[24] In this treatment, participants are explicitly informed that they are asked for their belief about "how likely it is that others would donate", thus making it clear that the question is about others' donation choices. For non-donors, while beliefs are not significantly different across *NonInc* and *Inc*, beliefs are significantly higher in *Inc-Karni-Exp* ($p < 0.01$), relative to *NonInc* (Figure 3). However, we do not find a significant difference between *Inc-Karni* and *Inc-Karni-Exp* ($p > 0.10$). These results are robust to the inclusion of demographic controls in regression analyses (Appendix C) and suggest that both the ability of the Karni mechanism to be structured as a payment question and the mechanism itself play a role in mitigating self-serving biases.

Figure 3: Beliefs by elicitation mechanism (including *Inc-Karni-Exp*) for donors and non-donors



Notes: Mann-Whitney test, error bars represent standard errors. Dotted line represents the empirical benchmark. *** denotes $p < 0.01$; ** denotes $p < 0.05$; * denotes $p < 0.10$; ns denotes $p > 0.10$.

Inconsistent switching in *Inc-Karni*: As explained in Section 3, we exclude approximately 23% of subjects from all *Inc-Karni* treatments as we are unable to identify their beliefs due to inconsistent switching behaviour. This inconsistent switching behaviour could be an indication of confusion, indifference, or non-standard preferences. Möbius et al. (2022) also report multiple switching in 13% to 22% of subjects. Dave et al. (2010) report similar findings using the Holt and Laury (2002) multiple price list procedure with the proportion of inconsistent choices ranging from 5% for subjects with higher math scores to more than 20% for subjects with lower math scores.

[24] From the $N = 128$, we excluded $N = 23$, or 18% of participants due to multiple switching.

As a robustness check, we create a proxy for multiple switchers' beliefs by summing the number of Option A choices, and find that these beliefs do not differ from that of single switchers.[25] Another alternative would have been to enforce a single switching point, however doing so would prevent us from identifying confusion, indifference or non-standard preferences and add more noise to the data. This highlights a potential limitation of the Karni mechanism as heterogeneity in cognitive ability, as indicated by Raven's scores, among subjects could affect the quality of data collected (Burfurd and Wilkening, 2021). We find no significant differences in cognitive ability between treatments for the full sample. However, once we exclude subjects with inconsistent switching behaviour, we find that the average Raven's score is significantly higher in *Inc-Karni* than *NonInc* (2.38 vs. 2.03, one-tailed MW test, $p = 0.01$). Note that in Tables 3 and 4 our main results hold even after controlling for cognitive ability. Given that previous work predicts *more* motivated reasoning from individuals with higher cognitive ability (Chen and Heese, 2021), having such a sample in *Inc-Karni* would be a bias *against* our results. Despite having a sample with slightly higher cognitive ability, our finding that belief distortions are less likely in *Inc-Karni* thus strengthens our main result.

Cognitive uncertainty: A related explanation for the beliefs in *Inc-Karni* is that cognitive uncertainty causes participants to revert to simple heuristics such as the 50% or midpoint default (e.g., Enke and Graeber, 2021). Schlag and Tremewan (2021) observe a more frequent belief of 50% when using the Karni mechanism compared to their "frequency method" and that this belief is more likely in subjects with low scores in the Cognitive Reflection Test (CRT).[26] We find no clear pattern between subjects' Ravens scores and beliefs in *Inc-Karni* (see Appendix D). We conducted an additional treatment, *Inc-Karni-Survey* ($N = 51$), as a robustness check of *Inc-Karni* with survey questions about subjects' decision-making processes. When asked about how they made their switching decision, more than 90% of participants indicated, in open-ended responses, that they considered the likelihood that a previous participant chose to donate, with a majority of these subjects being single switchers. This suggests that participants understood that their earnings would be maximised by switching close to their belief about the subjective probability, as opposed to reverting to a cognitive default due to confusion.[27]

Framing effects: Another possibility is that framing effects contributed to the different beliefs across the two incentivised treatments. Critcher and Dunning (2013) find that beliefs elicited (without an incentive) using an 'individual frame', i.e., regarding a single other, are higher than those elicited using a 'population frame', i.e., regarding the whole population. Bauer and Wolff (2018) argue that a population frame strengthens the consensus effect in a strategic setting. In our

[25] Similar approaches can be found in Holt and Laury (2002) and Bandyopadhyay et al. (2021).

[26] The frequency method is similar to the question in *NonInc* and *Inc*, though developed independently.

[27] An example of a response was: "I thought about the odds and at what point it was worth it to choose option B and how reasonable my chances were and if I could trust other participants."

experiment, *Inc-Karni* has a stronger individual frame (although the framing used in *NonInc* and *Inc* lies somewhere in between an individual and a population frame) and we find that the beliefs of non-donors are higher in *Inc-Karni* than the other two treatments. However, if our result in *Inc-Karni* is indeed driven by a framing effect, then we should similarly observe lower beliefs by donors, who should be equally affected by framing. Since this is not the case, we can conclude that framing alone is not driving our main results.

Taken together, our finding of more accurate beliefs in *Inc-Karni* cannot be fully explained by the exclusion of inconsistent switchers, cognitive uncertainty, nor by framing effects. Instead, our hypothesis that self-serving concerns are less salient while monetary incentives are more salient in *Inc-Karni* remains the most likely explanation to organise our data.

6.2 Why do beliefs differ between donors and non-donors?

Section 6.2 examines potential explanations for the belief gap between donors and non-donors under introspection and a simple incentive. We first investigate whether the timing of the donation decision and belief elicitation affects the belief response and report results from three additional treatments (*NoAsk*, *NoAsk-Inc-Exp* and *Inc-Ask-Rev*) which suggest that non-donors do not distort their beliefs directly in response to a single donation ask, but rather are consistent in holding biased beliefs about others. We then consider the (false) consensus effect and individual differences in optimism levels as possible alternative explanations for the belief gap.

Timing of belief elicitation Given that we find evidence of biased beliefs in non-donors, we explore whether the donation ask in our experiment causes a distortion of beliefs, or whether beliefs are robust to the timing of belief elicitation, such that we capture underlying types of agents using our donor/non-donor classification. We conducted an additional treatment, *NoAsk*, for each of the three mechanisms, *NoAsk-NonInc* ($N = 91$), *NoAsk-Inc* ($N = 101$) and *NoAsk-Inc-Karni* ($N = 133$), in which participants are not asked to make a personal donation.[28] Similar to the original treatments *Ask*, subjects are asked to choose a charity (to control for any priming effects) and report their beliefs about the proportion of previous donors. Overall, we find no significant difference in beliefs between *Ask* and *NoAsk* ($p > 0.10$, Table 5).[29] According to a Kolmogorov-Smirnov (KS) test, the distribution of beliefs between *Ask* and *NoAsk* is not significantly different ($p > 0.10$) for any of the three mechanisms.[30]

One possibility is that having selected a charity in Stage 1, participants anticipated an upcoming donation ask in *NoAsk* and adjusted their beliefs accordingly. We conducted an additional treatment, *NoAsk-Inc-Exp* ($N = 101$), in which subjects were explicitly informed that they will

[28] We excluded $N = 38$, or 29% of participants in *NoAsk-Inc-Karni* due to multiple switching.

[29] Ging-Jehli et al. (2020) also find that third-party beliefs do not differ from that of other players.

[30] These results are also confirmed by the Epps-Singleton test (Epps and Singleton, 1986).

not be asked to make a personal donation.[31] We find no difference between *NoAsk-Inc* and *NoAsk-Inc-Exp* in either mean beliefs (4.59 vs. 4.99, $p > 0.10$) or in the distribution of beliefs (KS test, $p > 0.10$), offering support that subjects did not anticipate a donation opportunity.

Table 5: Beliefs in the *Ask* and *NoAsk* treatments

	(1)	(2)
NoAsk	-0.20	-0.16
	(0.22)	(0.22)
Inc	0.54	0.54
	(0.27)	(0.28)
Inc-Karni	1.09^{***}	1.08^{***}
	(0.27)	(0.28)
Constant	4.52^{***}	2.67
	(0.22)	(1.70)
H_0: *Inc* = *Inc-Karni*	$p = 0.05$	$p = 0.06$
Controls	*No*	*Yes*
R^2	0.03	0.08
Adj. R^2	0.02	0.05
Num. obs.	598	598

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the subject's belief about the proportion of donors. The baseline Treatment is *NonInc*. Statistical significance accounts for multiple hypothesis testing (adjusted using the Bonferroni correction). The control variables are: gender, age, education, religiosity, political ideology, income and cognitive ability.

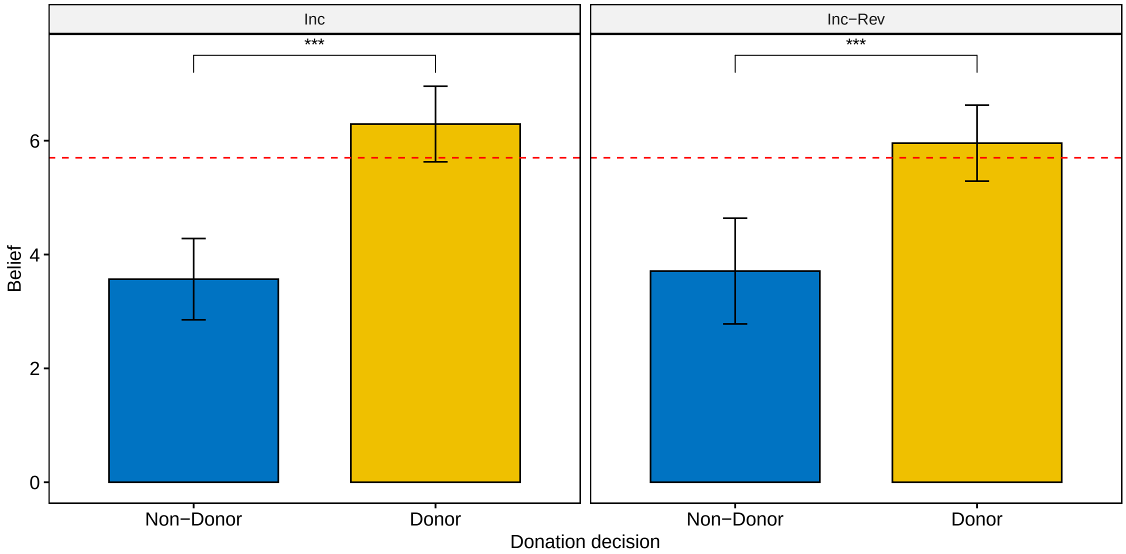
While we find substantial heterogeneity in beliefs for donors and non-donors in *Ask*, we are unable to identify this in *NoAsk* since we do not observe donation choices. We conducted an additional treatment, *Inc-Rev* ($N = 99$), in which we reverse the order of tasks from *Inc*, such that we first elicit incentivised beliefs about others, followed by a surprise donation decision. Figure 4 shows that donors' beliefs remain significantly higher than that of non-donors ($p < 0.01$). Non-donors report an average belief of 3.71, which is lower than the true proportion (Wilcoxon signed-rank test, $p < 0.01$) while donors report a belief of 5.96, which is not significantly different from the empirical benchmark (Wilcoxon signed-rank test, $p > 0.10$). Donation rates also do not differ based on the timing of the donation ask in *Inc* and *Inc-Rev* (64% vs. 69%, χ^2 test, $p > 0.10$).

Similar to previous work (e.g., Ging-Jehli et al., 2020), our results show that the opportunity to donate *per se* does not cause a distortion in beliefs about others, rather the belief biases we observe in non-donors persist, irrespective of when belief elicitation occurs.[32] By manipulating the timing

[31] We chose to run the additional treatments with *Inc* because participants have an incentive to think carefully about their decisions while self-serving concerns still appear to be relevant under this mechanism.

[32] This contrasts with previous papers which manipulate the timing of information provided to participants about a potential self-serving motive (e.g., Babcock and Loewenstein, 1995; Gneezy et al., 2020; Saccardo and Serra-Garcia, 2022; Bicchieri et al., 2020) and find that the timing matters for beliefs.

Figure 4: Beliefs of donors and non-donors in *Inc* and *Inc-Rev*



Notes: Mann-Whitney test, error bars represent standard errors. Dotted line represents the empirical benchmark. *** denotes $p < 0.01$; ** denotes $p < 0.05$; * denotes $p < 0.10$; ns denotes $p > 0.10$.

of belief elicitation, we observe the existence of a non-giving type not only in behaviour (de Oliveira et al., 2011), but also in beliefs. When given the opportunity, these agents consistently choose not to donate, and are also consistent in holding biased beliefs about others' behaviour. One possible explanation is that subjects are likely to have encountered numerous donation solicitations in their lifetime. For non-giving types, this means that their beliefs may have already been distorted by previous experiences.

The (false) consensus effect A potential alternative explanation for the belief gap between donors and non-donors is the (false) consensus effect, whereby people believe others are generally similar to themselves and project their own "type" onto others.[33] Evidence of a consensus bias has been found in both psychology (e.g., Ross et al., 1977) and economics (e.g., Selten and Ockenfels, 1998; Bicchieri and Xiao, 2009; Breitmoser, 2019; Erkal et al., 2021). In the context of our experiment, a pure projection bias would predict that non-donors underestimate the proportion of donors, while donors should overestimate the donation rate (i.e., $\hat{\lambda} > \lambda$). We do not observe this in our data. Instead, our results show that donors' beliefs are accurate, and that what *appears* to be a consensus effect is in fact driven by more selfish types. This is consistent with our theoretical framework, in which donors have no incentive to incur psychological costs to distort their beliefs, but for non-donors the gains in self-image potentially exceed these costs. Iriberri and Rey-Biel (2013) also report that while selfish types believe that 87% of others would choose the same action

[33] Engelmann and Strobel (2000) argue that a consensus effect is only 'false' if individuals attach greater weights to their own decisions than that of a randomly selected individual from the population.

that they chose, more prosocial types report a belief that is closer to 50%. Further, even if we suppose that a consensus effect is contributing in part to the belief gap in *NonInc*, it is unable to explain the difference *between* the incentivised mechanisms, i.e., the Karni mechanism results in significantly higher beliefs in non-donors, without having any effect on donors' beliefs.

Optimism To investigate the possibility that the belief gap between donors and non-donors is driven by levels of optimism (as an individual trait), we included an additional survey question in *Inc-NoAsk-Exp* and *Inc-Rev*, asking for self-reported optimism.[34] We do not find any evidence that non-donors are more pessimistic than donors ($p > 0.10$) in a general context.

In sum, we show that the belief gap between donors and non-donors in *NonInc* and *Inc* does not depend on the timing of belief elicitation as biases persist even when this timing is reversed. We further argue that our results are not driven by a pure consensus effect as this would also predict a positive bias in donors, which is not consistent with the data. We rule out individual levels of optimism as a major driver of the belief gap based on survey data showing that self-reported optimism is not higher in donors than non-donors.

7 Conclusion

Growing evidence points to the importance of beliefs in explaining behaviour that preferences alone are unable to explain. Based on a simple theoretical framework which captures the tension between utility derived from monetary payoffs and self-image, we design an experiment involving the opportunity to donate to charity and compare three commonly used methods (non-incentivised, incentivised and Karni) to elicit beliefs about giving behaviour. We investigate whether participants who choose not to give are more likely to hold biased beliefs about others under introspection and whether beliefs vary with the elicitation mechanism.

Our key takeaways can be summarised as follows: First, when belief accuracy is not incentivised, individuals with weaker altruistic motivations are more likely to reveal beliefs that are biased by self-serving concerns. These belief distortions are robust to the timing of belief elicitation and point to the existence of giving and non-giving types in both behaviour and beliefs. Our results support the provision of accurate information to encourage prosocial behaviour (e.g., Shang and Croson, 2009), and offer a potential explanation for why this may work in organisations, i.e., calibrating the beliefs of non-giving types can help to restrict belief distortions and increase the costs of maintaining a positive self-image, thus encouraging more prosocial behaviour.

Second, introducing a simple incentive is not sufficient in reducing biases in non-donors' beliefs. However, these beliefs become substantially more accurate under the more complex Karni

[34] The following question was asked: "On the following scale (where 1 = not optimistic at all and 10 = extremely optimistic) how optimistic do you consider yourself to be?"

mechanism, despite the monetary cost of reporting an inaccurate belief being lower in *Inc-Karni* than in *Inc*. This is consistent with the idea that monetary payoffs are made more salient while self-serving concerns are made less salient in *Inc-Karni*. We therefore caution that different elicitation mechanisms can produce different results. The elicitation mechanism used should depend on whether belief biases are the focus of the research question, or whether the goal is to minimise these biases to allow other effects to surface. For the former, survey methods which directly ask for beliefs may be sufficient, while adding a simple incentive can be useful in encouraging more careful introspection. Regarding the latter, merely introducing incentives may not be enough and researchers should consider using more complex mechanisms such as Karni to "de-motivate" beliefs.

An important open question is which method provides the best approximation of "true" beliefs, i.e., the beliefs that feed into decision making. If the ultimate goal is to identify the beliefs that map into decisions, more complex mechanisms may be less suitable, if certain motivations are amplified in a way that is inconsistent with the actual decision-making environment. Our findings suggest that the belief biases of non-giving types are robust to the timing of belief elicitation. A promising avenue for future work is to examine the direction of this causality, namely do individuals act selfishly because they are better able to distort their beliefs to justify their actions, or do these biased beliefs come from underlying social preferences? Another interesting question is whether other aspects of belief elicitation might enhance or limit belief distortion, such as the incentive stake size or publicising beliefs.

References

Andreoni, J. (1989). Giving with impure altruism : Applications to charity and Ricardian equivalence. *Journal of Political Economy*, 97(6):1447–1458.

Andreoni, J. and Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753.

Andreoni, J. and Sanchez, A. (2020). Fooling myself or fooling observers? Avoiding social pressures by manipulating perceptions of deservingness of others. *Economic Inquiry*, 58(1):12–33.

Andreoni, J. and Serra-Garcia, M. (2021). Time inconsistent charitable giving. *Journal of Public Economics*, 198:104391.

Babcock, L. and Loewenstein, G. (1995). Biased judgments of fairness in bargaining. *The American Economic Review*, 85(5):1337–1343.

Baillon, A., Bleichrodt, H., and Granic, G. D. (2022). Incentives in surveys. *Journal of Economic Psychology*, 93:102552.

Bandyopadhyay, A., Begum, L., and Grossman, P. J. (2021). Gender differences in the stability of risk attitudes. *Journal of Risk and Uncertainty*, 63(2):169–201.

Bartling, B. and Özdemir, Y. (2022). The limits to moral erosion in markets: Social norms and the replacement excuse. *Available at SSRN 3043728*.

Bauer, D. and Wolff, I. (2018). Biases in beliefs: Experimental evidence. *TWI Research Paper Series, 109*.

Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–164.

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bicchieri, C., Dimant, E., and Sonderegger, S. (2020). It's Not a Lie If You Believe the Norm Does Not Apply: Conditional Norm-Following with Strategic Beliefs. *Working Paper*, (January):1–59.

Bicchieri, C., Dimant, E., and Xiao, E. (2021). Deviant or wrong? The effects of norm information on the efficacy of punishment. *Journal of Economic Behavior and Organization*, 188:209–235.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208.

Bodner, R. and Prelec, D. (2003). The Diagnostic Value of Actions in a Self-Signaling Model. *The Psychology of Economic Decisions, Volume I: Rationality and Well-Being*, 1:105–126.

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

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

Breitmoser, Y. (2019). Knowing me, imagining you: Projection and overbidding in auctions. *Games and Economic Behavior*, 113:423–447.

Bullock, J. G., Gerber, A. S., Hill, S. J., and Huber, G. A. (2013). Partisan bias in factual beliefs about politics. *National Bureau of Economic Research.*

Burfurd, I. and Wilkening, T. (2021). Cognitive heterogeneity and complex belief elicitation. *Experimental Economics*, pages 1–36. https://doi.org/10.1007/s10683-021-09722-x.

Carpenter, J. (2021). The shape of warm glow: Field experimental evidence from a fundraiser. *Journal of Economic Behavior and Organization*, 191:555–574.

Charness, G., Gneezy, U., and Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior & Organization*, 189:234–256.

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

Chen, S. and Heese, C. (2021). Fishing for good news: Motivated information acquisition.

Coutts, A. (2019). Testing models of belief bias: An experiment. *Games and Economic Behavior*, 113:549–565.

Critcher, C. R. and Dunning, D. (2013). Predicting persons' versus a person's goodness: Behavioral forecasts diverge for individuals versus populations. *Journal of Personality and Social Psychology*, 104(1):28–44.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Danilov, A., Khalmetski, K., and Sliwka, D. (2021). Descriptive norms and guilt aversion. *Journal of Economic Behavior & Organization*, 191:293–311.

Danz, D., Vesterlund, L., and Wilson, A. J. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review.*

Dave, C., Eckel, C. C., Johnson, C. A., and Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, 41(3):219–243.

de Oliveira, A. C., Croson, R. T., and Eckel, C. (2011). The giving type: Identifying donors. *Journal of Public Economics*, 95(5-6):428–435.

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015). Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism. *American Economic Review*, 105(11):3416–3442.

Dimant, E. and Gesche, T. (2021). Nudging enforcers: How norm perceptions and motives for lying shape sanctions.

Drobner, C. (2022). Motivated beliefs and anticipation of uncertainty resolution. *American Economic Review: Insights*, 4(1):89–105.

Ducharme, W. M. and Donnell, M. L. (1973). Intrasubject comparison of four response modes for "subjective probability" assessment. *Organizational Behavior and Human Performance*, 10(1):108–117.

Engelmann, D. and Strobel, M. (2000). The false consensus effect disappears if representative information and monetary incentives are given. *Experimental Economics*, 3(3):241–260.

Enke, B., Gneezy, U., Hall, B., Martin, D., Nelidov, V., Offerman, T., and van de Ven, J. (2021). Cognitive biases: Mistakes or missing stakes? *The Review of Economics and Statistics*, pages 1–45. https://doi.org/10.1162/rest_a_01093.

Enke, B. and Graeber, T. (2021). Cognitive uncertainty. *National Bureau of Economic Research (No. w26518).*

Epley, N. and Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives*, 30(3):133–140.

Epps, T. and Singleton, K. J. (1986). An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*, 26(3-4):177–203.

Erkal, N., Gangadharan, L., and Koh, B. H. (2020). Replication: Belief elicitation with quadratic and binarized scoring rules. *Journal of Economic Psychology*, 81:102315.

Erkal, N., Gangadharan, L., and Koh, B. H. (2021). By chance or by choice? Biased attribution of others' outcomes when social preferences matter. *Experimental Economics*, pages 1–31. https://doi.org/10.1007/s10683-021-09731-w.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *Review of Economic Studies*, 83(2):587–628.

Exley, C. L. and Petrie, R. (2018). The impact of a surprise donation ask. *Journal of Public Economics*, 158:152–167.

Freeman, D. J. and Mayraz, G. (2019). Why choice lists increase risk taking. *Experimental Economics*, 22:131–154.

Gabaix, X. (2019). Behavioral inattention. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, pages 261–343. Elsevier.

Gangadharan, L., Grossman, P. J., Jones, K., and Leister, C. M. (2018). Paternalistic giving: Restricting recipient choice. *Journal of Economic Behavior and Organization*, 151:143–170.

Gangadharan, L., Grossman, P. J., and Xue, N. (2023). Using willingness to pay to measure the strength of altruistic motives. *Economics Letters*, 226:111073.

Ging-Jehli, N. R., Schneider, F. H., and Weber, R. A. (2020). On self-serving strategic beliefs. *Games and Economic Behavior*, 122:341–353.

Gino, F., Norton, M. I., and Weber, R. A. (2016). Motivated Bayesians: Feeling moral while acting egoistically. *Journal of Economic Perspectives*, 30(3):189–212.

Gneezy, U., Saccardo, S., Serra-Garcia, M., and van Veldhuizen, R. (2020). Bribing the Self. *Games and Economic Behavior*, 120:311–324.

Grossman, Z. and van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1):173–217.

Haisley, E. C. and Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior*, 68(2):614–625.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., and Bigham, J. P. (2018). A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655.

Holt, C. A. and Smith, A. M. (2016). Belief elicitation with a synchronized lottery choice menu that is invariant to risk attitudes. *American Economic Journal: Microeconomics*, 8(1):110–139.

Horton, J. J., Rand, D. G., and Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425.

Hossain, T. and Okui, R. (2013). The binarized scoring rule. *Review of Economic Studies*, 80(3):984–1001.

Iriberri, N. and Rey-Biel, P. (2013). Elicited beliefs and social information in modified dictator games: What do dictators believe other dictators do? *Quantitative Economics*, 4(3):515–547.

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica*, 77(2):603–606.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, 108(3):480.

List, J. A. and Gallet, C. A. (2001). What experimental protocol influence disparities between actual and hypothetical stated values? *Environmental and Resource Economics*, 20(3):241–254.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science*.

Molnár, A. and Heintz, C. (2016). Beliefs about people's prosociality: Eliciting predictions in dictator games. *Department of Economics - CEU working papers series*, (January).

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

Null, C. (2011). Warm glow, information, and inefficient charitable giving. *Journal of Public Economics*, 95(5-6):455–465.

Ottoni-Wilhelm, M., Vesterlund, L., and Xie, H. (2017). Why do people give? Testing pure and impure altruism. *American Economic Review*, 107(11):3617–33.

Raven, J. C. and Court, J. (1938). *Raven's Progressive Matrices*. Western Psychological Services Los Angeles, CA.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3):279–301.

Saccardo, S. and Serra-Garcia, M. (2022). Enabling or limiting cognitive flexibility? evidence of demand for moral commitment. *American Economic Review*.

Schlag, K. and Tremewan, J. (2021). Simple belief elicitation: An experimental evaluation. *Journal of Risk and Uncertainty*, 62(2):137–155.

Schlag, K. H., Tremewan, J., and van der Weele, J. J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, 18(3):457–490.

Schotter, A. and Trevino, I. (2014). Belief Elicitation in the Laboratory. *Annual Review of Economics*, 6(1):103–128.

Schwardmann, P., Tripodi, E., and Van der Weele, J. J. (2022). Self-persuasion: Evidence from field experiments at international debating competitions. *American Economic Review*, 112(4):1118–46.

Selten, R. and Ockenfels, A. (1998). An experimental solidarity game. *Journal of Economic Behavior and Organization*, 34(4):517–539.

Serra-Garcia, M. and Szech, N. (2021). The (in) elasticity of moral ignorance. *Management Science*. https://doi.org/10.1287/mnsc.2021.4153.

Shang, J. and Croson, R. (2009). A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *The Economic Journal*, 119(540):1422–1439.

Snowberg, E. and Yariv, L. (2021). Testing the waters: Behavior across participant pools. *American Economic Review*, 111(2):687–719.

Solda, A., Ke, C., Page, L., and Von Hippel, W. (2020). Strategically delusional. *Experimental Economics*, 23(3):604–631.

Taylor, S. E. and Thompson, S. C. (1982). Stalking the elusive "vividness" effect. *Psychological Review*, 89(2):155.

Tonin, M. and Vlassopoulos, M. (2013). Experimental evidence of self-image concerns as motivation for giving. *Journal of Economic Behavior and Organization*, 90:19–27.

Trautmann, S. T. and van de Kuilen, G. (2015). Belief elicitation: A horse race among truth serums. *The Economic Journal*, 125(589):2116–2135.

Valero, V. (2021). Redistribution and beliefs about the source of income inequality. *Experimental Economics*, pages 1–26. https://doi.org/10.1007/s10683-021-09733-8.

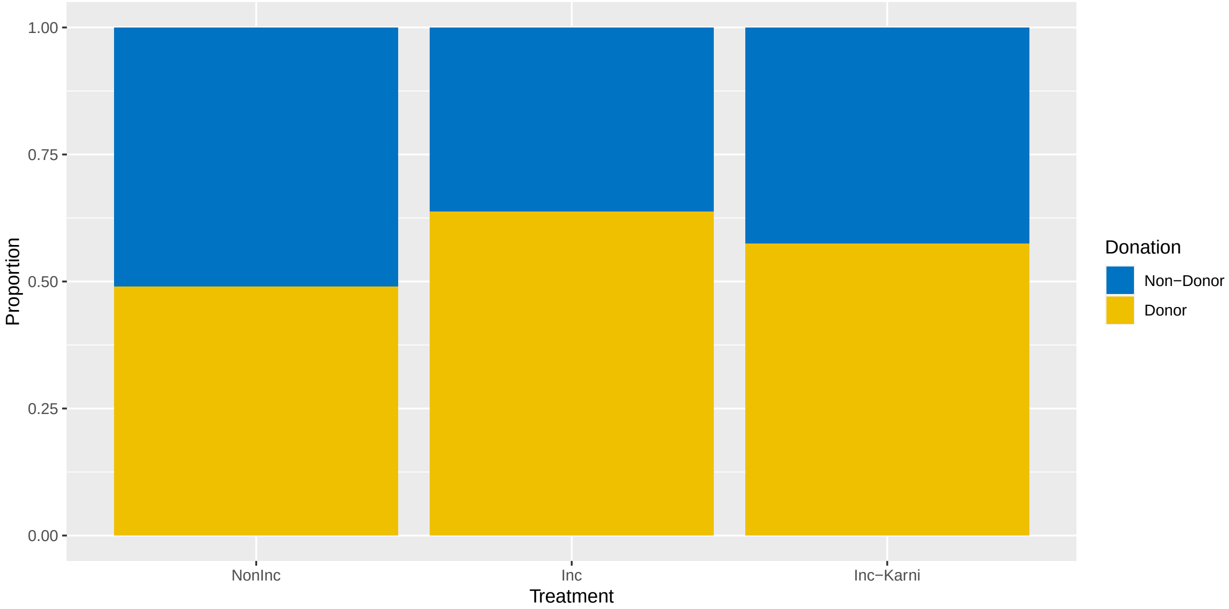
Van der Weele, J. J., Kulisa, J., Kosfeld, M., and Friebel, G. (2014). Resisting moral wiggle room: how robust is reciprocal behavior? *American Economic Journal: Microeconomics*, 6(3):256–64.

Zimmermann, F. (2020). The dynamics of motivated beliefs. *American Economic Review*, 110(2):337–363.

A Donation rates

Figure A.1 presents the proportion of donors and non-donors in *NonInc*, *Inc* and *Inc-Karni*. According to a χ^2 test, the donation rates are not significantly different across treatments at the 5% level ($p = 0.07$).

Figure A.1: Donation rates



Note: Error bars represent standard errors.

B Regression results excluding implemented donations

Table B.1: Beliefs in *NonInc* (excluding implemented donations)

	(1)	(2)
Non-Donor	-3.13^{***}	-3.11^{***}
	(0.49)	(0.49)
Raven's score		0.15
		(0.17)
Constant	6.07^{***}	5.95^{***}
	(0.36)	(1.35)
Controls	*No*	*Yes*
R^2	0.30	0.52
Adj. R^2	0.30	0.40
Num. obs.	94	94

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the subject's belief about the proportion of donors. Participants whose donations were implemented in Stage 1 are excluded. Statistical significance accounts for multiple hypothesis testing (adjusted using the Bonferroni correction). The control variables are: gender, age, education, religiosity, political ideology and income.

Table B.2: Beliefs of donors and non-donors (excluding implemented donations)

	Non-Donors		Donors		Pooled	
	(1)	(2)	(3)	(4)	(5)	(6)
Inc	0.63	0.45	0.11	0.15	0.11	-0.03
	(0.52)	(0.58)	(0.52)	(0.58)	(0.50)	(0.54)
Inc-Karni	2.19^{***}	2.08^{***}	-0.40	-0.31	-0.40	-0.44
	(0.46)	(0.50)	(0.53)	(0.61)	(0.52)	(0.56)
Raven's score		-0.13		0.14		-0.02
		(0.15)		(0.16)		(0.11)
Non-Donor					-3.13^{***}	-3.01^{***}
					(0.51)	(0.54)
Inc x Non-Donor					0.52	0.59
					(0.73)	(0.78)
Inc-Karni x Non-Donor					2.59^{***}	2.63^{***}
					(0.71)	(0.73)
Constant	2.94^{***}	2.74	6.07^{***}	4.40^{**}	6.07^{***}	5.27^{***}
	(0.33)	(1.23)	(0.39)	(1.46)	(0.38)	(0.98)
H_0: *Inc* = *Inc-Karni*	$p < 0.01$	$p < 0.01$	$p = 0.31$	$p = 0.42$	$p = 0.29$	$p = 0.43$
Controls	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
R^2	0.14	0.26	0.01	0.11	0.20	0.23
Adj. R^2	0.13	0.13	-0.01	-0.04	0.19	0.16
Num. obs.	143	143	150	150	293	293

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the subject's belief about the proportion of donors. The baseline Treatment is *NonInc*. Participants whose donations were implemented in Stage 1 are excluded. Statistical significance accounts for multiple hypothesis testing (adjusted using the Bonferroni correction). The control variables are: gender, age, education, religiosity, political ideology and income.

C Beliefs of donors and non-donors in ***Inc-Karni-Exp***

Table C.1: Beliefs of donors and non-donors (including *Inc-Karni-Exp*)

	Donors		Non-Donors		Pooled	
	(1)	(2)	(3)	(4)	(5)	(6)
Inc	0.63	0.62	0.46	0.45	0.46	0.32
	(0.51)	(0.55)	(0.47)	(0.51)	(0.46)	(0.48)
Inc-Karni	2.19^{***}	2.24^{***}	-0.11	-0.03	-0.11	-0.05
	(0.46)	(0.48)	(0.48)	(0.52)	(0.47)	(0.49)
Inc-Karni-Exp	1.29^{**}	1.41^{**}	-0.24	-0.06	-0.24	-0.06
	(0.46)	(0.48)	(0.50)	(0.54)	(0.49)	(0.51)
Raven's score		-0.09		0.03		-0.04
		(0.12)		(0.12)		(0.08)
Non-Donor					-2.96^{***}	-2.85^{***}
					(0.49)	(0.50)
Inc x Non-Donor					0.16	0.35
					(0.70)	(0.72)
Inc-Karni x Non-Donor					2.29^{***}	2.33^{***}
					(0.67)	(0.68)
Inc-Karni-Exp x Non-Donor					1.54^{*}	1.30
					(0.68)	(0.70)
Constant	2.94^{***}	2.95^{**}	5.90^{***}	4.26^{***}	5.90^{***}	4.94^{***}
	(0.33)	(0.99)	(0.36)	(1.09)	(0.35)	(0.78)
H_0: *Inc* = *Inc-Karni*	$p < 0.01$	$p < 0.01$	$p = 0.21$	$p = 0.32$	$p = 0.20$	$p = 0.43$
H_0: *Inc* = *Inc-Karni-Exp*	$p = 0.19$	$p = 0.15$	$p = 0.14$	$p = 0.33$	$p = 0.13$	$p = 0.44$
H_0: *Inc-Karni* = *Inc-Karni-Exp*	$p = 0.05$	$p = 0.08$	$p = 0.78$	$p = 0.96$	$p = 0.77$	$p = 0.98$
Controls	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
R^2	0.11	0.21	0.01	0.10	0.18	0.22
Adj. R^2	0.10	0.12	-0.00	-0.00	0.17	0.17
Num. obs.	199	199	219	219	418	418

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the subject's belief about average generosity. The baseline Treatment is *NonInc*. Statistical significance accounts for multiple hypothesis testing (adjusted using the Bonferroni correction). The control variables are: gender, age, education, religiosity and income.

Table C.2: Beliefs of donors and non-donors (including *Inc-Karni-Exp*)

	Non-Donors		Donors		Pooled	
	(1)	(2)	(3)	(4)	(5)	(6)
NonInc	-1.29^{**}	-1.41^{**}	0.24	0.06	0.24	0.06
	(0.46)	(0.48)	(0.50)	(0.54)	(0.49)	(0.51)
Inc	-0.66	-0.80	0.71	0.50	0.71	0.38
	(0.50)	(0.55)	(0.47)	(0.51)	(0.46)	(0.48)
Inc-Karni	0.90	0.83	0.14	0.02	0.14	0.01
	(0.45)	(0.47)	(0.48)	(0.51)	(0.47)	(0.49)
Raven's score		-0.09		0.03		-0.04
		(0.12)		(0.12)		(0.08)
Non-Donor					-1.42^{**}	-1.56^{***}
					(0.48)	(0.49)
NonInc x Non-Donor					-1.54^{*}	-1.30
					(0.68)	(0.70)
Inc x Non-Donor					-1.37	-0.95
					(0.69)	(0.72)
Inc-Karni x Non-Donor					0.76	1.04
					(0.66)	(0.67)
Constant	4.23^{***}	4.36^{***}	5.65^{***}	4.20^{***}	5.65^{***}	4.88^{***}
	(0.32)	(1.03)	(0.36)	(1.05)	(0.35)	(0.76)
Controls	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
R^2	0.11	0.21	0.01	0.10	0.18	0.22
Adj. R^2	0.10	0.12	-0.00	-0.00	0.17	0.17
Num. obs.	199	199	219	219	418	418

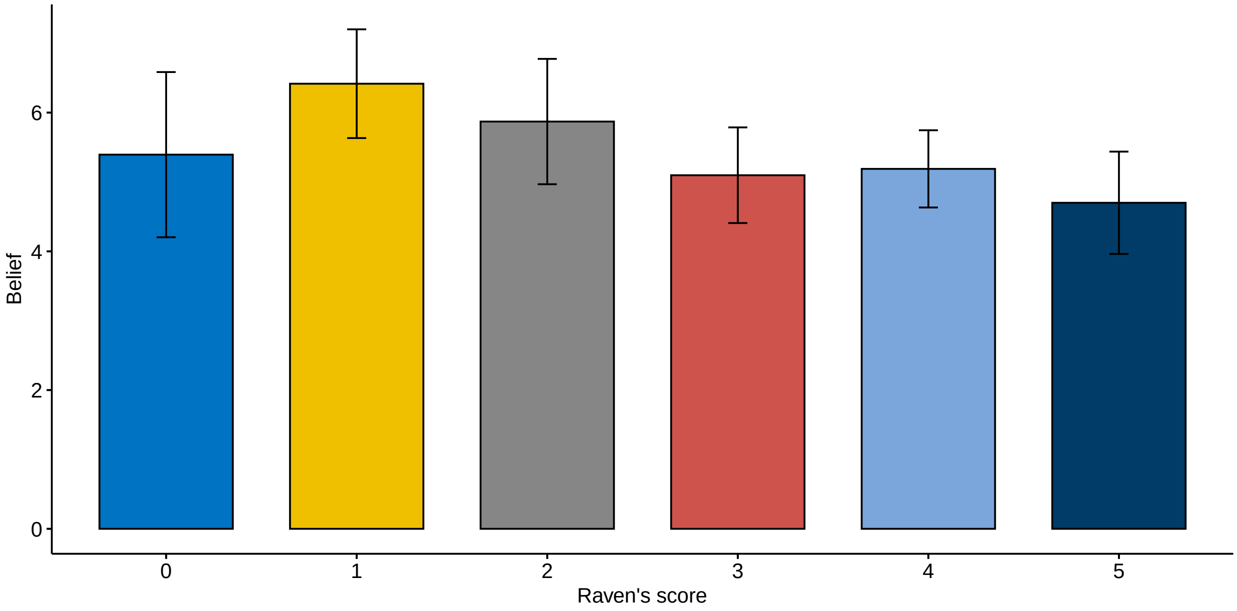
$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the subject's belief about average generosity. The baseline Treatment is *Inc-Karni-Exp*. Statistical significance accounts for multiple hypothesis testing (adjusted using the Bonferroni correction). The control variables are: gender, age, education, religiosity and income.

D Cognitive ability

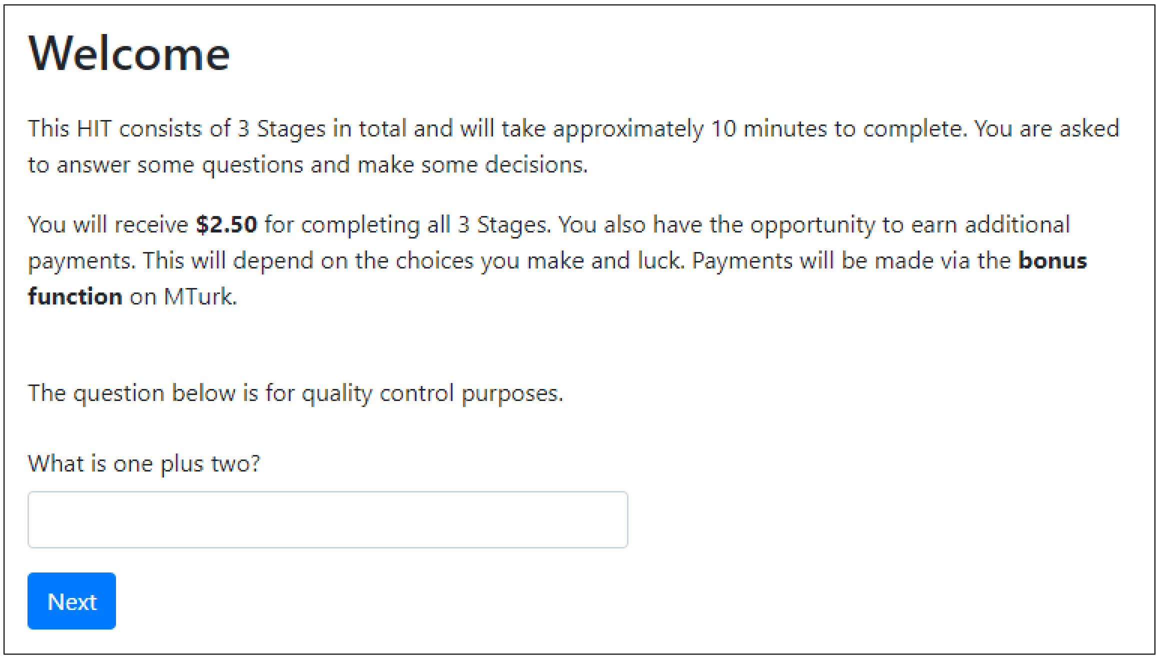
We investigate the relationship between cognitive ability and beliefs in *Inc-Karni* and find no clear pattern in beliefs (Figure D.1). According to a Pearson correlation test, the correlation coefficient between Raven's scores and beliefs in *Inc-Karni* is -0.18 ($p = 0.06$). However, we find no clear relationship between Raven's scores and the likelihood of switching in the middle (i.e., reporting a belief of 5.5). According to a Wilcoxon signed-rank test, beliefs are different from 5.5 only when Raven's score is equal to 1 ($p = 0.09$).

Figure D.1: Beliefs by Raven's score in *Inc-Karni*



Note: Error bars represent standard errors.

E Instructions



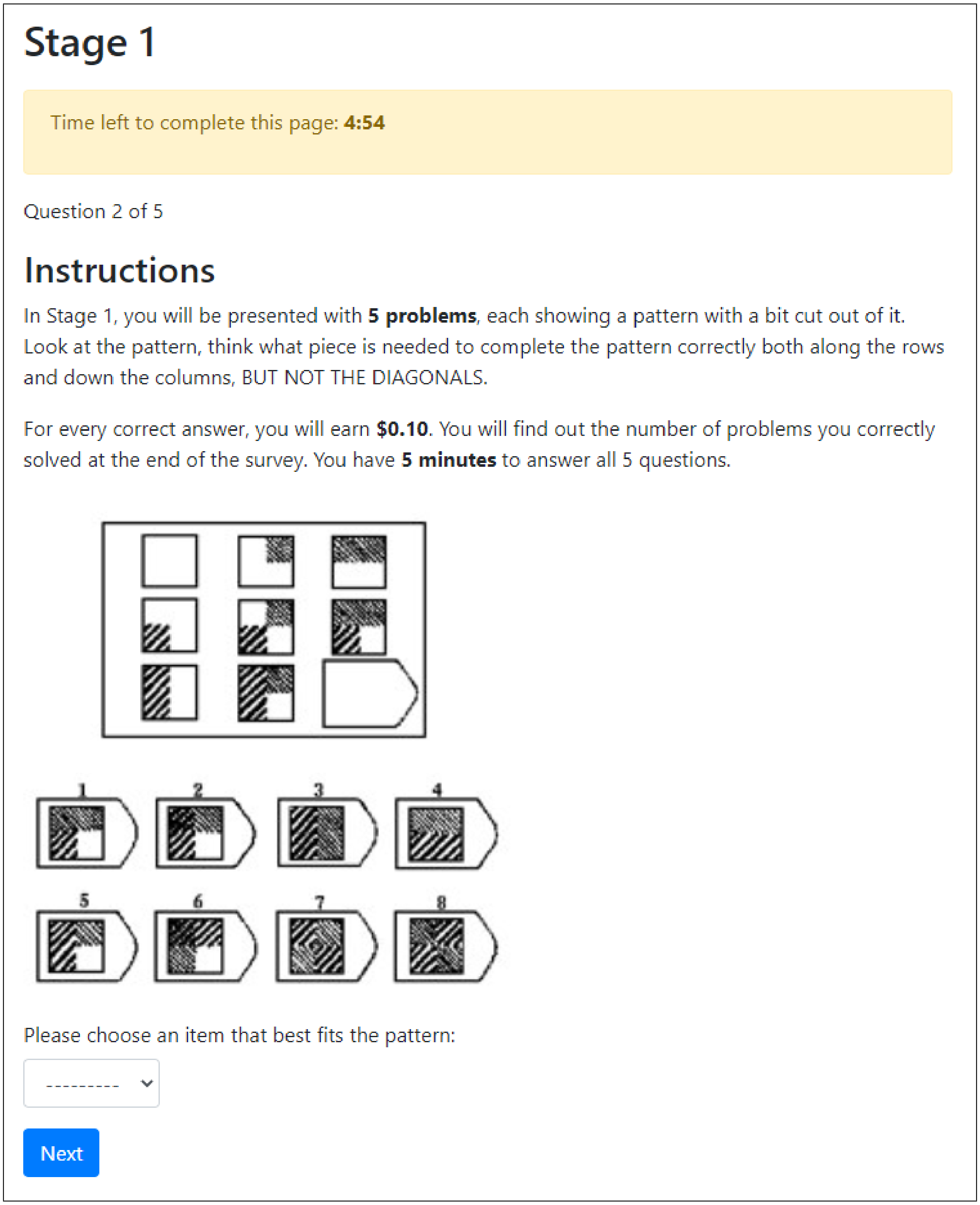
Welcome

This HIT consists of 3 Stages in total and will take approximately 10 minutes to complete. You are asked to answer some questions and make some decisions.

You will receive **$2.50** for completing all 3 Stages. You also have the opportunity to earn additional payments. This will depend on the choices you make and luck. Payments will be made via the **bonus function** on MTurk.

The question below is for quality control purposes.

What is one plus two?



Stage 1

Time left to complete this page: **4:54**

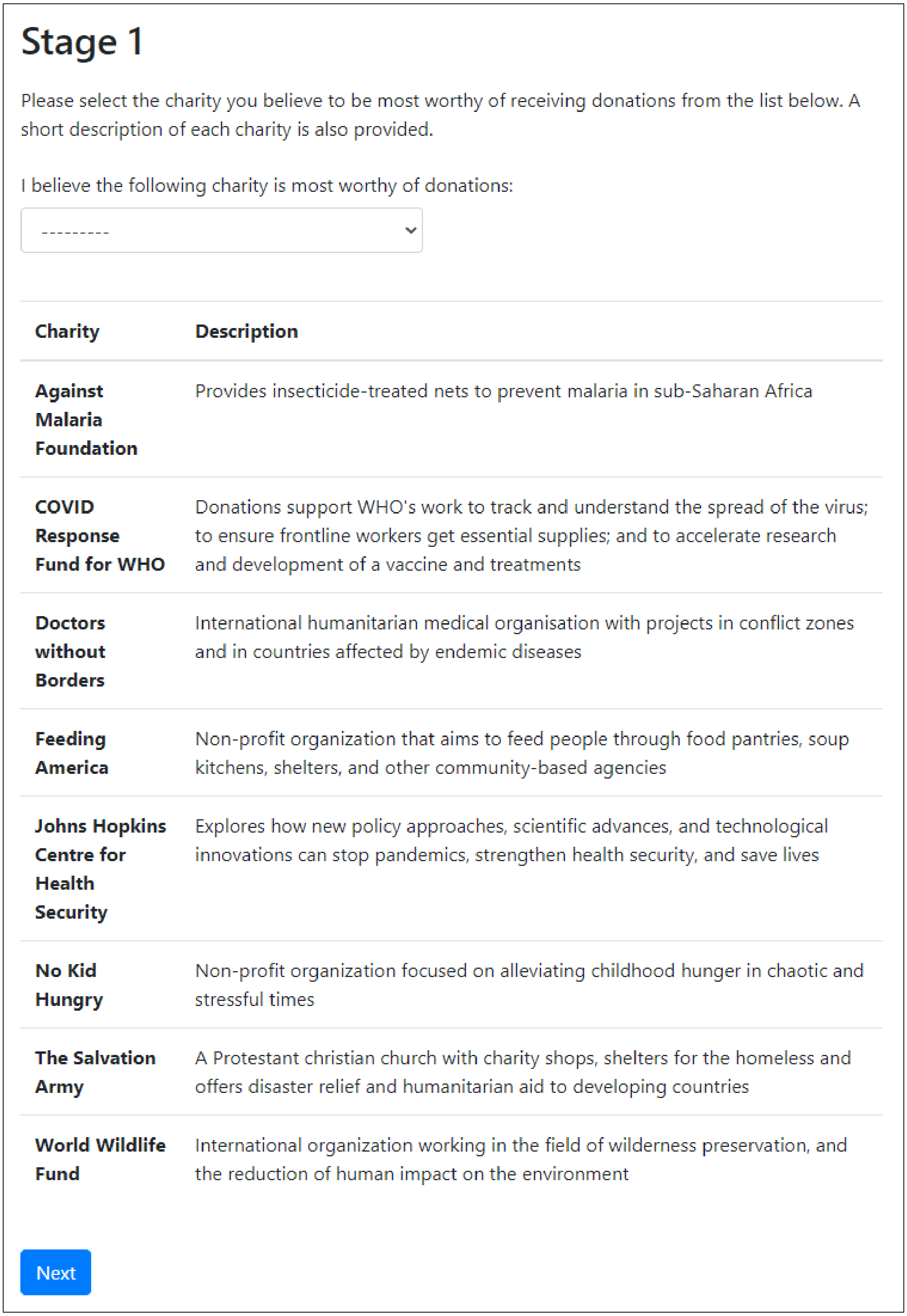
Question 2 of 5

Instructions

In Stage 1, you will be presented with **5 problems**, each showing a pattern with a bit cut out of it. Look at the pattern, think what piece is needed to complete the pattern correctly both along the rows and down the columns, BUT NOT THE DIAGONALS.

For every correct answer, you will earn **$0.10**. You will find out the number of problems you correctly solved at the end of the survey. You have **5 minutes** to answer all 5 questions.

Please choose an item that best fits the pattern:

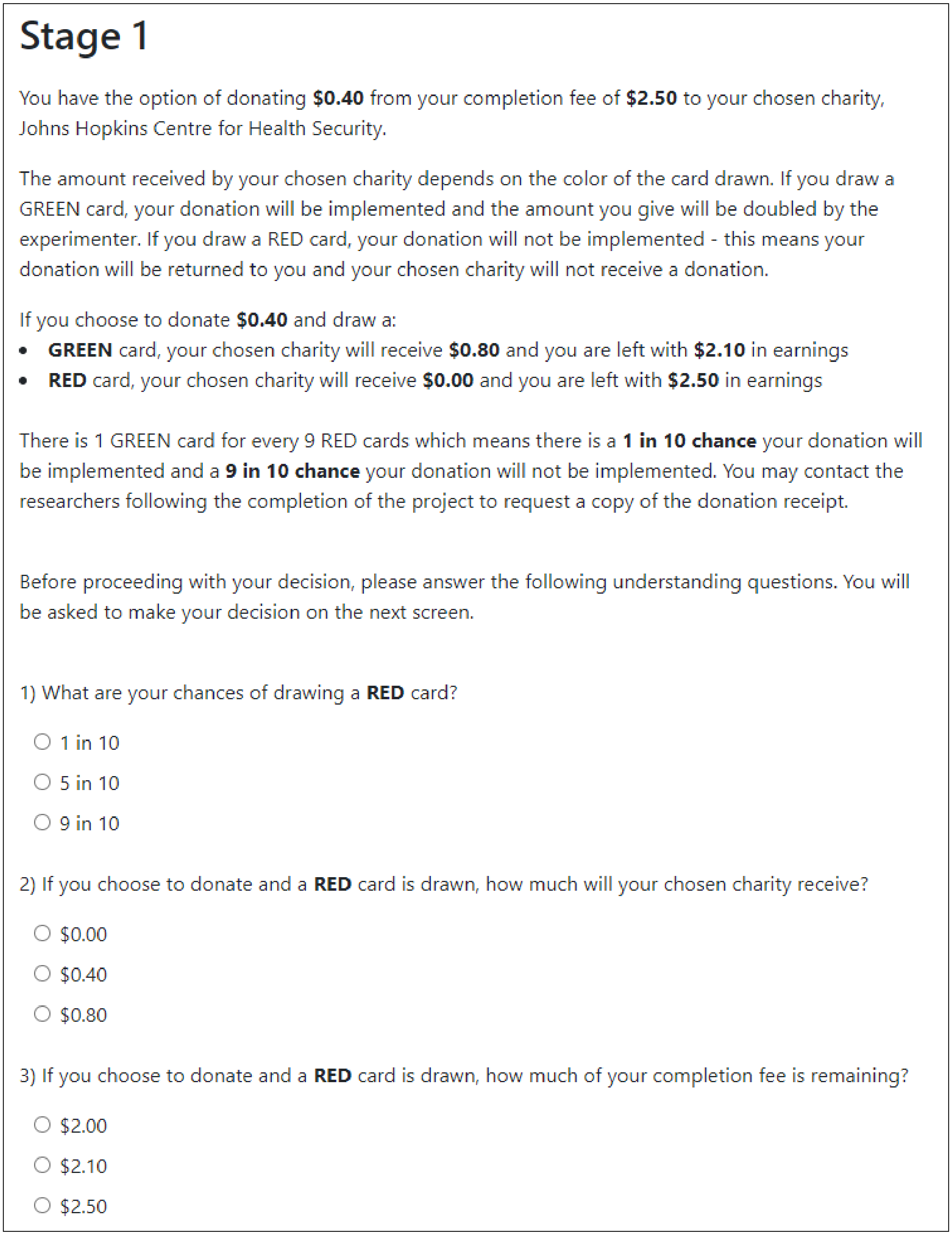


Stage 1

Please select the charity you believe to be most worthy of receiving donations from the list below. A short description of each charity is also provided.

I believe the following charity is most worthy of donations:

Charity	Description
Against Malaria Foundation	Provides insecticide-treated nets to prevent malaria in sub-Saharan Africa
COVID Response Fund for WHO	Donations support WHO's work to track and understand the spread of the virus; to ensure frontline workers get essential supplies; and to accelerate research and development of a vaccine and treatments
Doctors without Borders	International humanitarian medical organisation with projects in conflict zones and in countries affected by endemic diseases
Feeding America	Non-profit organization that aims to feed people through food pantries, soup kitchens, shelters, and other community-based agencies
Johns Hopkins Centre for Health Security	Explores how new policy approaches, scientific advances, and technological innovations can stop pandemics, strengthen health security, and save lives
No Kid Hungry	Non-profit organization focused on alleviating childhood hunger in chaotic and stressful times
The Salvation Army	A Protestant christian church with charity shops, shelters for the homeless and offers disaster relief and humanitarian aid to developing countries
World Wildlife Fund	International organization working in the field of wilderness preservation, and the reduction of human impact on the environment



Stage 1

You have the option of donating **$0.40** from your completion fee of **$2.50** to your chosen charity, Johns Hopkins Centre for Health Security.

The amount received by your chosen charity depends on the color of the card drawn. If you draw a GREEN card, your donation will be implemented and the amount you give will be doubled by the experimenter. If you draw a RED card, your donation will not be implemented - this means your donation will be returned to you and your chosen charity will not receive a donation.

If you choose to donate **$0.40** and draw a:

- **GREEN** card, your chosen charity will receive **$0.80** and you are left with **$2.10** in earnings
- **RED** card, your chosen charity will receive **$0.00** and you are left with **$2.50** in earnings

There is 1 GREEN card for every 9 RED cards which means there is a **1 in 10 chance** your donation will be implemented and a **9 in 10 chance** your donation will not be implemented. You may contact the researchers following the completion of the project to request a copy of the donation receipt.

Before proceeding with your decision, please answer the following understanding questions. You will be asked to make your decision on the next screen.

1) What are your chances of drawing a **RED** card?

○ 1 in 10

○ 5 in 10

○ 9 in 10

2) If you choose to donate and a **RED** card is drawn, how much will your chosen charity receive?

○ $0.00

○ $0.40

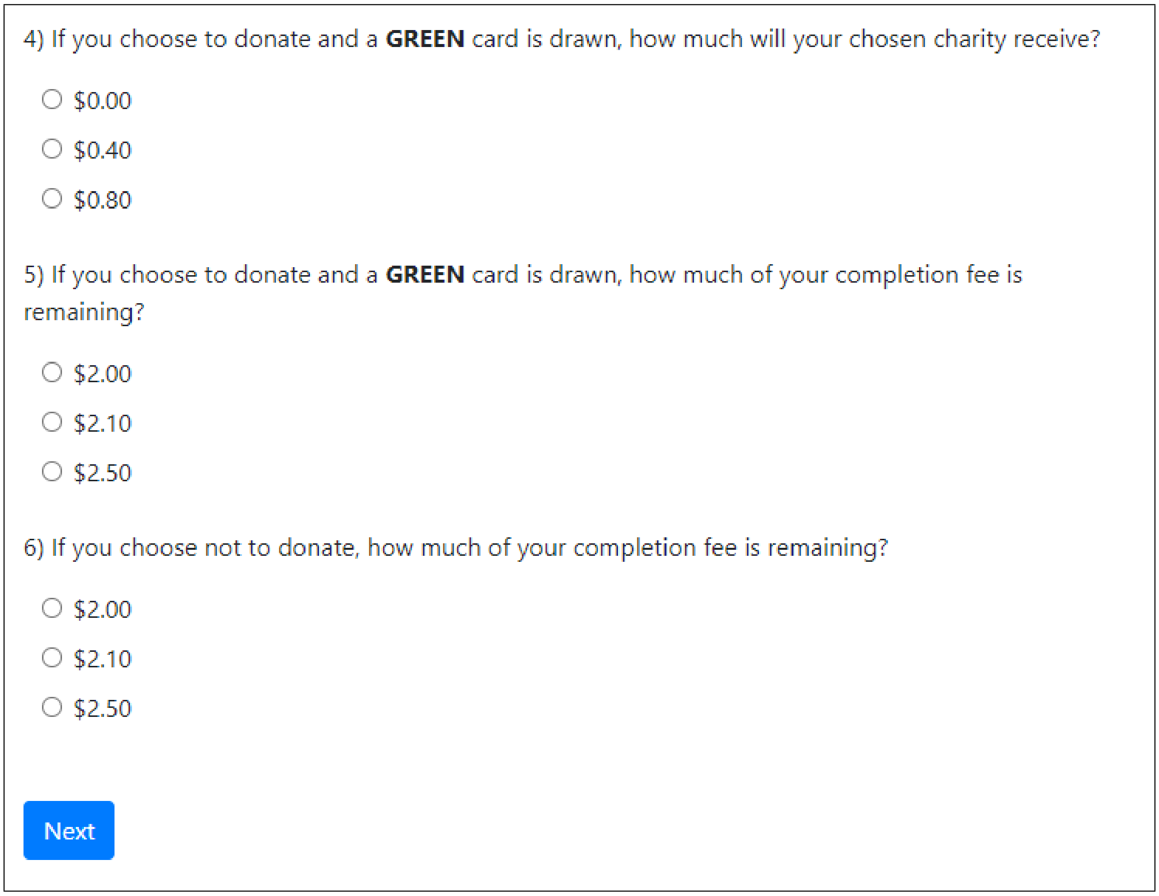
○ $0.80

3) If you choose to donate and a **RED** card is drawn, how much of your completion fee is remaining?

○ $2.00

○ $2.10

○ $2.50



4) If you choose to donate and a **GREEN** card is drawn, how much will your chosen charity receive?

○ $0.00

○ $0.40

○ $0.80

5) If you choose to donate and a **GREEN** card is drawn, how much of your completion fee is remaining?

○ $2.00

○ $2.10

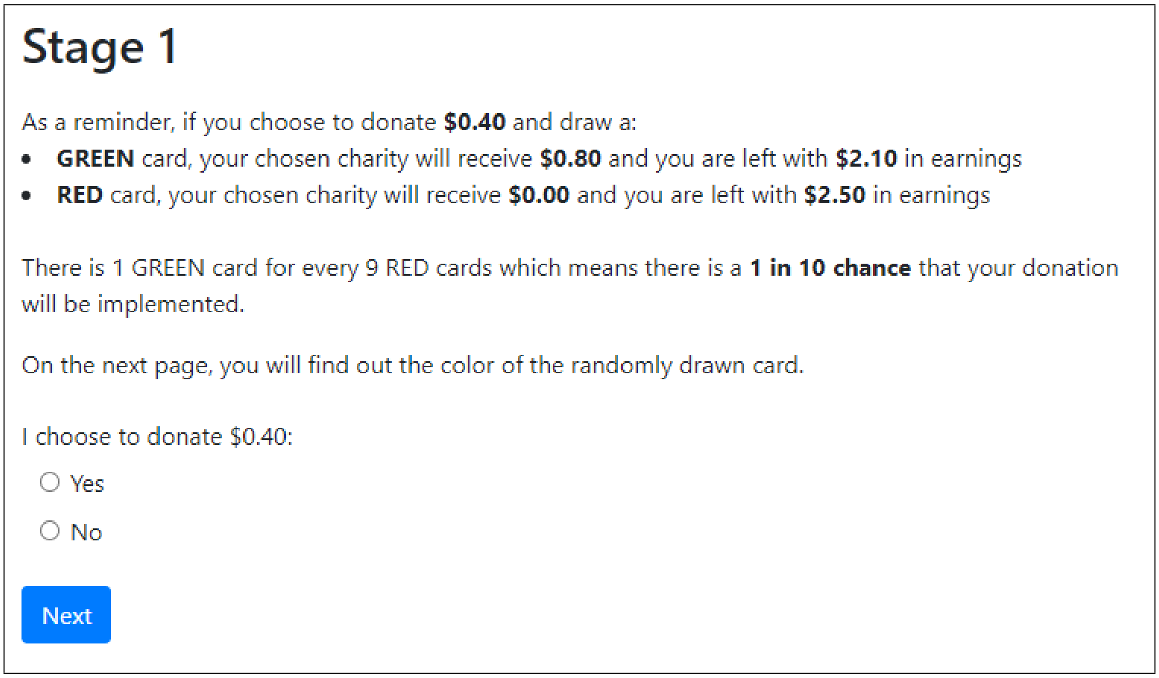
○ $2.50

6) If you choose not to donate, how much of your completion fee is remaining?

○ $2.00

○ $2.10

○ $2.50



Stage 1

As a reminder, if you choose to donate **$0.40** and draw a:

- **GREEN** card, your chosen charity will receive **$0.80** and you are left with **$2.10** in earnings
- **RED** card, your chosen charity will receive **$0.00** and you are left with **$2.50** in earnings

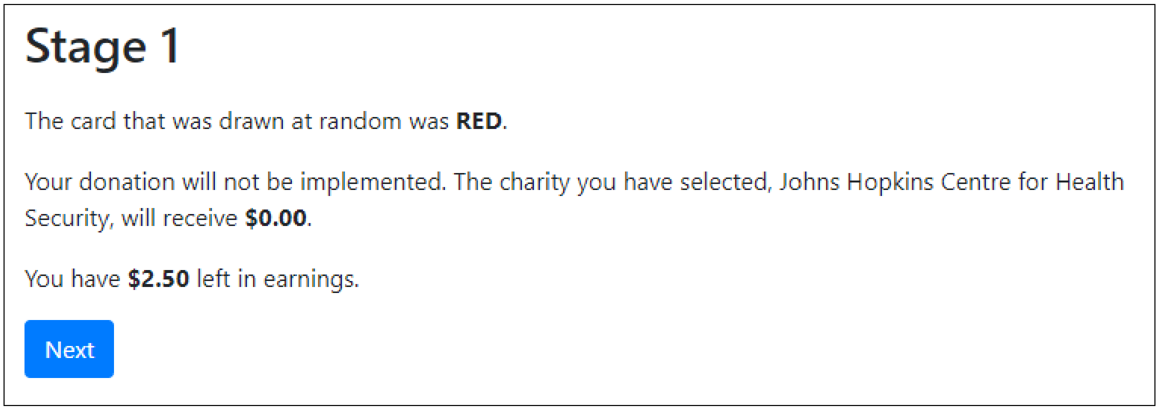
There is 1 GREEN card for every 9 RED cards which means there is a **1 in 10 chance** that your donation will be implemented.

On the next page, you will find out the color of the randomly drawn card.

I choose to donate $0.40:

○ Yes

○ No



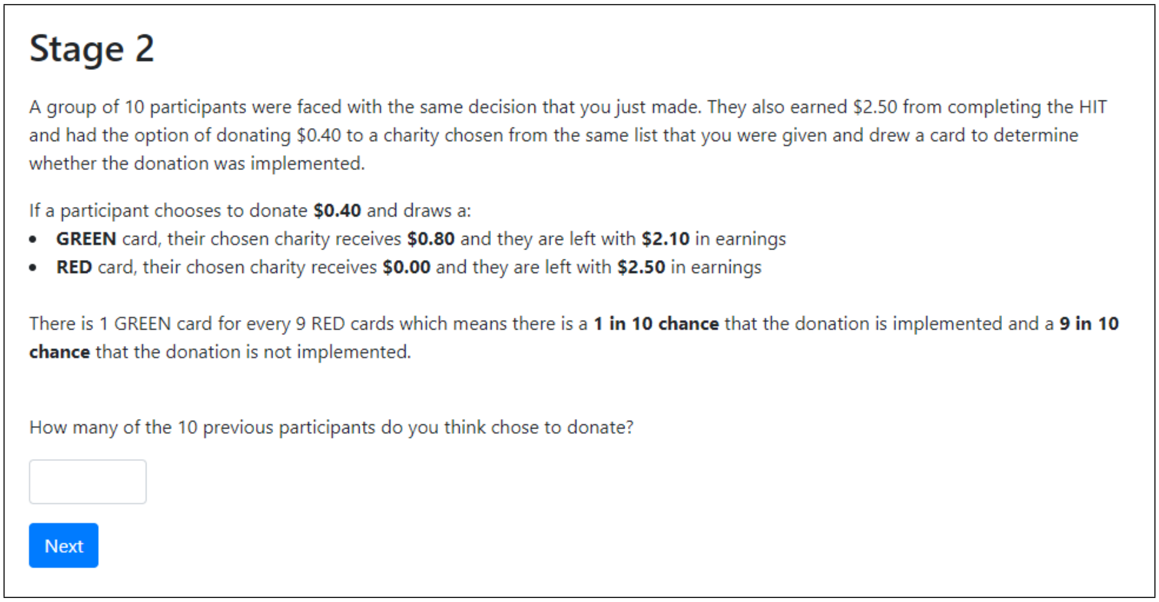
Stage 1

The card that was drawn at random was **RED**.

Your donation will not be implemented. The charity you have selected, Johns Hopkins Centre for Health Security, will receive **$0.00**.

You have **$2.50** left in earnings.

NonInc



Stage 2

A group of 10 participants were faced with the same decision that you just made. They also earned $2.50 from completing the HIT and had the option of donating $0.40 to a charity chosen from the same list that you were given and drew a card to determine whether the donation was implemented.

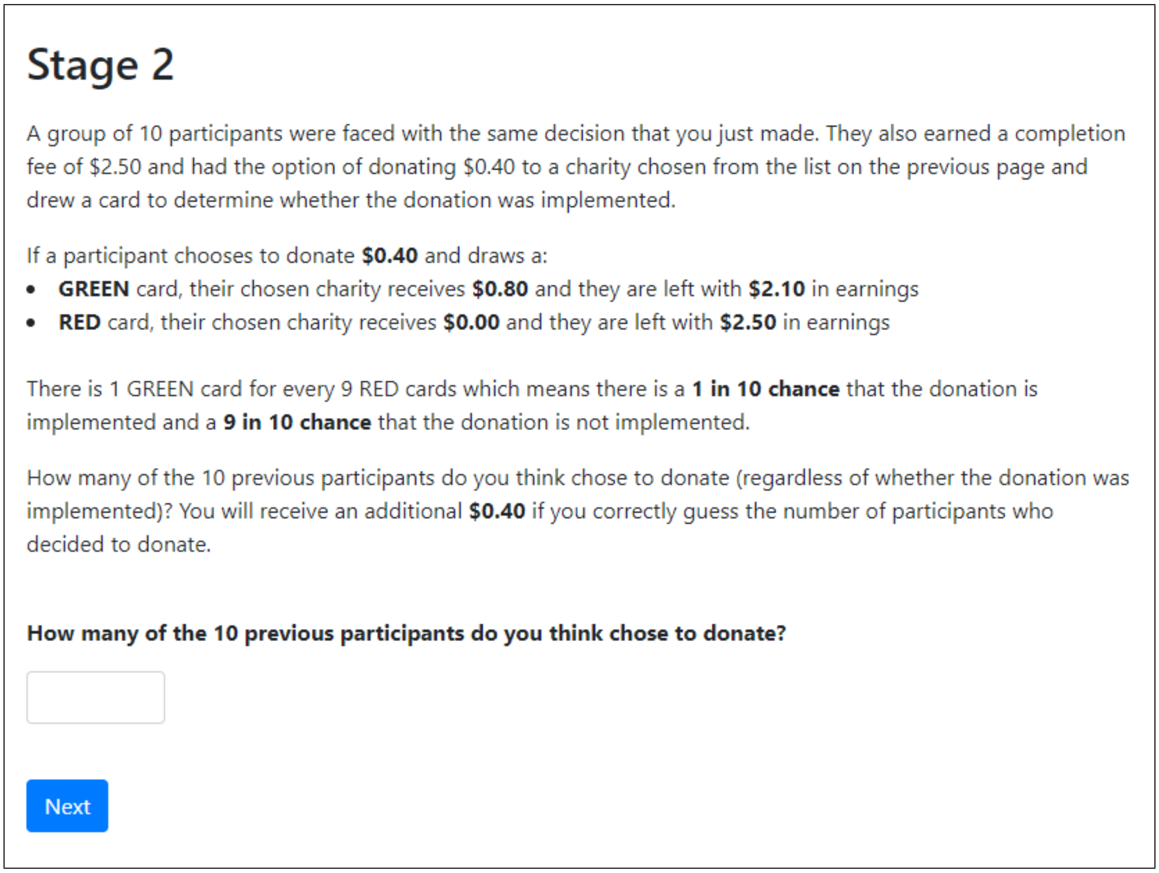
If a participant chooses to donate **$0.40** and draws a:

- **GREEN** card, their chosen charity receives **$0.80** and they are left with **$2.10** in earnings
- **RED** card, their chosen charity receives **$0.00** and they are left with **$2.50** in earnings

There is 1 GREEN card for every 9 RED cards which means there is a **1 in 10 chance** that the donation is implemented and a **9 in 10 chance** that the donation is not implemented.

How many of the 10 previous participants do you think chose to donate?

Inc



Stage 2

A group of 10 participants were faced with the same decision that you just made. They also earned a completion fee of $2.50 and had the option of donating $0.40 to a charity chosen from the list on the previous page and drew a card to determine whether the donation was implemented.

If a participant chooses to donate **$0.40** and draws a:

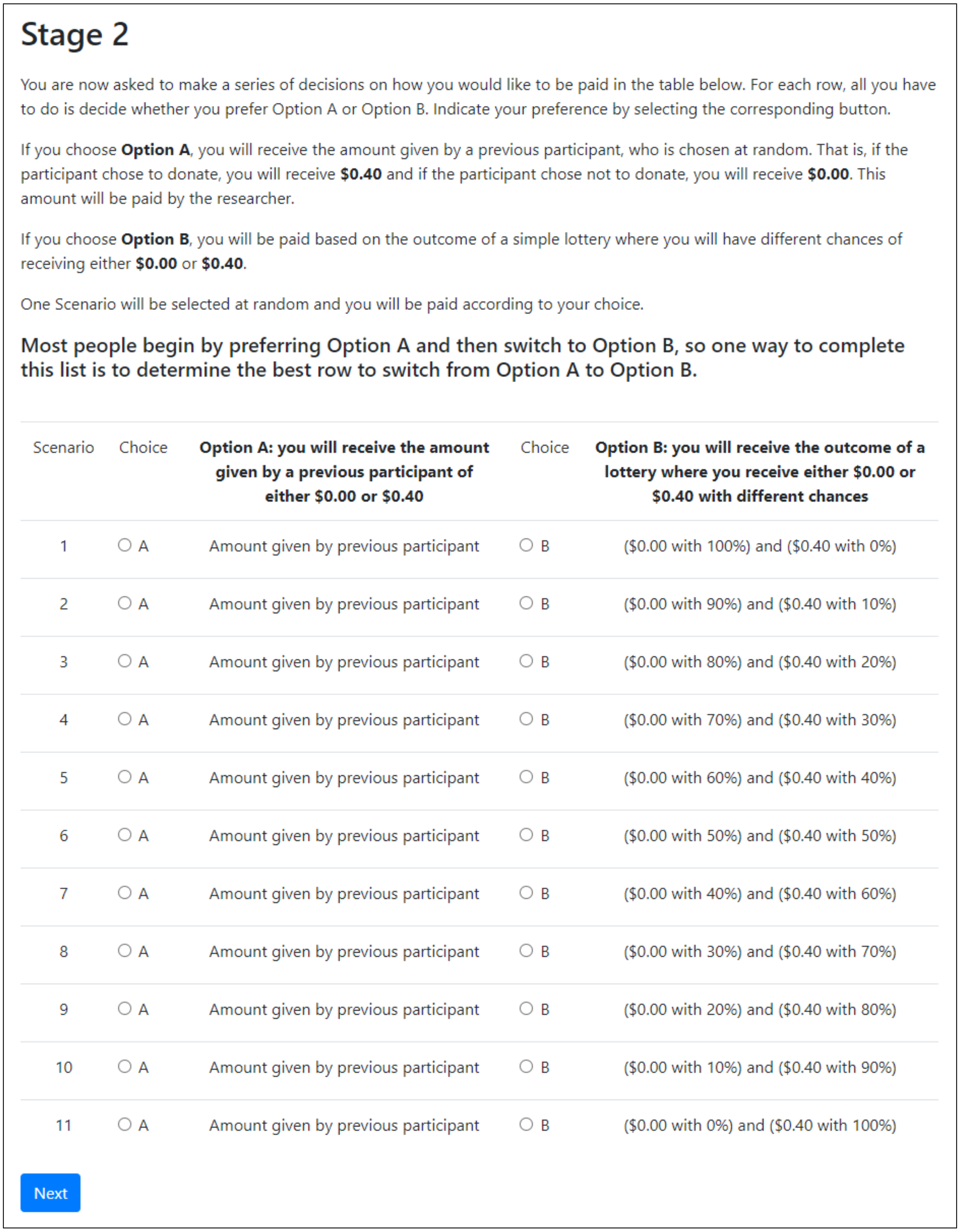
- **GREEN** card, their chosen charity receives **$0.80** and they are left with **$2.10** in earnings
- **RED** card, their chosen charity receives **$0.00** and they are left with **$2.50** in earnings

There is 1 GREEN card for every 9 RED cards which means there is a **1 in 10 chance** that the donation is implemented and a **9 in 10 chance** that the donation is not implemented.

How many of the 10 previous participants do you think chose to donate (regardless of whether the donation was implemented)? You will receive an additional **$0.40** if you correctly guess the number of participants who decided to donate.

How many of the 10 previous participants do you think chose to donate?

Inc-Karni



Stage 2

You are now asked to make a series of decisions on how you would like to be paid in the table below. For each row, all you have to do is decide whether you prefer Option A or Option B. Indicate your preference by selecting the corresponding button.

If you choose **Option A**, you will receive the amount given by a previous participant, who is chosen at random. That is, if the participant chose to donate, you will receive **$0.40** and if the participant chose not to donate, you will receive **$0.00**. This amount will be paid by the researcher.

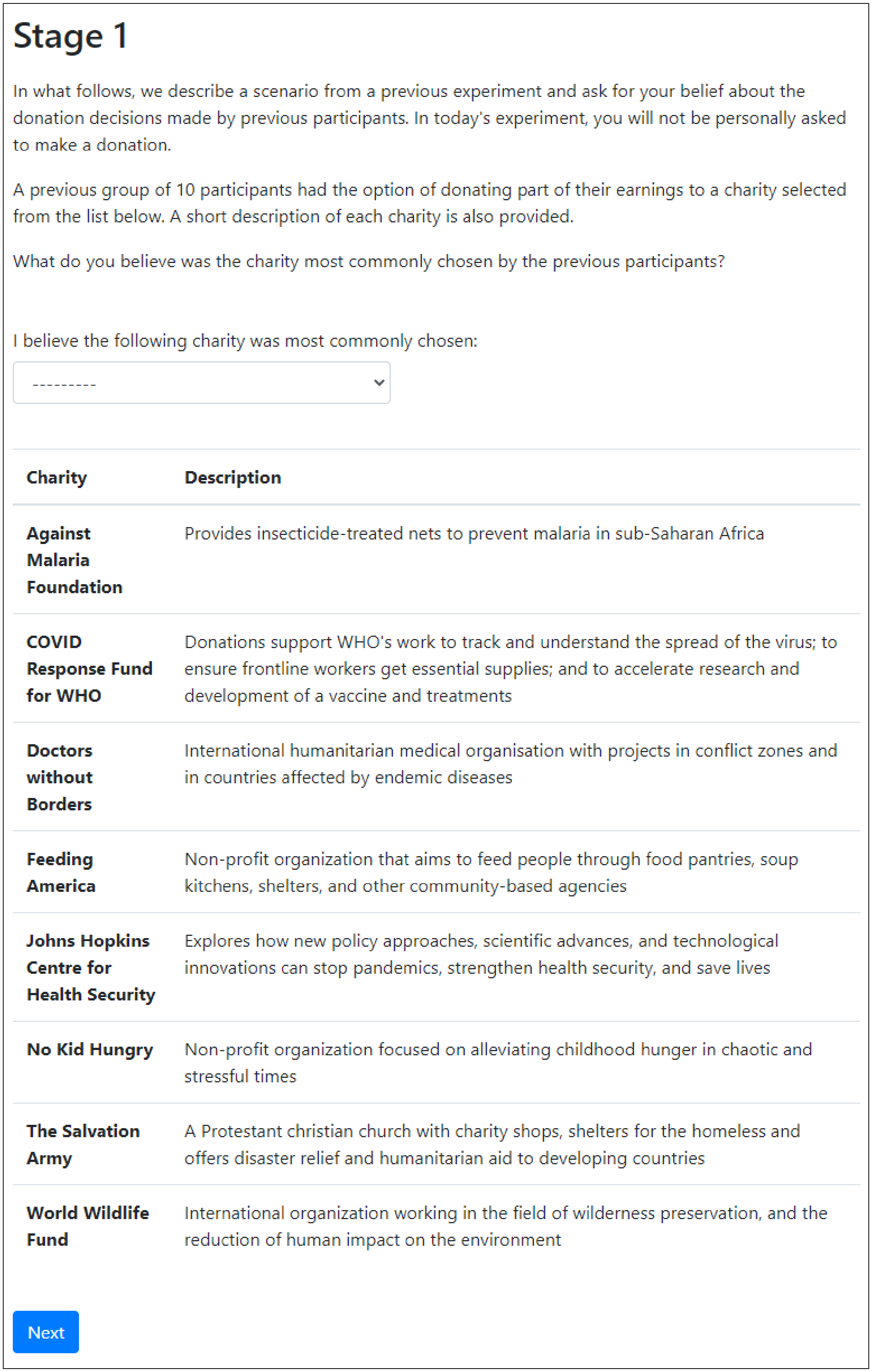
If you choose **Option B**, you will be paid based on the outcome of a simple lottery where you will have different chances of receiving either **$0.00** or **$0.40**.

One Scenario will be selected at random and you will be paid according to your choice.

Most people begin by preferring Option A and then switch to Option B, so one way to complete this list is to determine the best row to switch from Option A to Option B.

Scenario	Choice	**Option A: you will receive the amount given by a previous participant of either $0.00 or $0.40**	Choice	**Option B: you will receive the outcome of a lottery where you receive either $0.00 or $0.40 with different chances**
1	○ A	Amount given by previous participant	○ B	($0.00 with 100%) and ($0.40 with 0%)
2	○ A	Amount given by previous participant	○ B	($0.00 with 90%) and ($0.40 with 10%)
3	○ A	Amount given by previous participant	○ B	($0.00 with 80%) and ($0.40 with 20%)
4	○ A	Amount given by previous participant	○ B	($0.00 with 70%) and ($0.40 with 30%)
5	○ A	Amount given by previous participant	○ B	($0.00 with 60%) and ($0.40 with 40%)
6	○ A	Amount given by previous participant	○ B	($0.00 with 50%) and ($0.40 with 50%)
7	○ A	Amount given by previous participant	○ B	($0.00 with 40%) and ($0.40 with 60%)
8	○ A	Amount given by previous participant	○ B	($0.00 with 30%) and ($0.40 with 70%)
9	○ A	Amount given by previous participant	○ B	($0.00 with 20%) and ($0.40 with 80%)
10	○ A	Amount given by previous participant	○ B	($0.00 with 10%) and ($0.40 with 90%)
11	○ A	Amount given by previous participant	○ B	($0.00 with 0%) and ($0.40 with 100%)

Inc-NoAsk-Exp



Stage 1

In what follows, we describe a scenario from a previous experiment and ask for your belief about the donation decisions made by previous participants. In today's experiment, you will not be personally asked to make a donation.

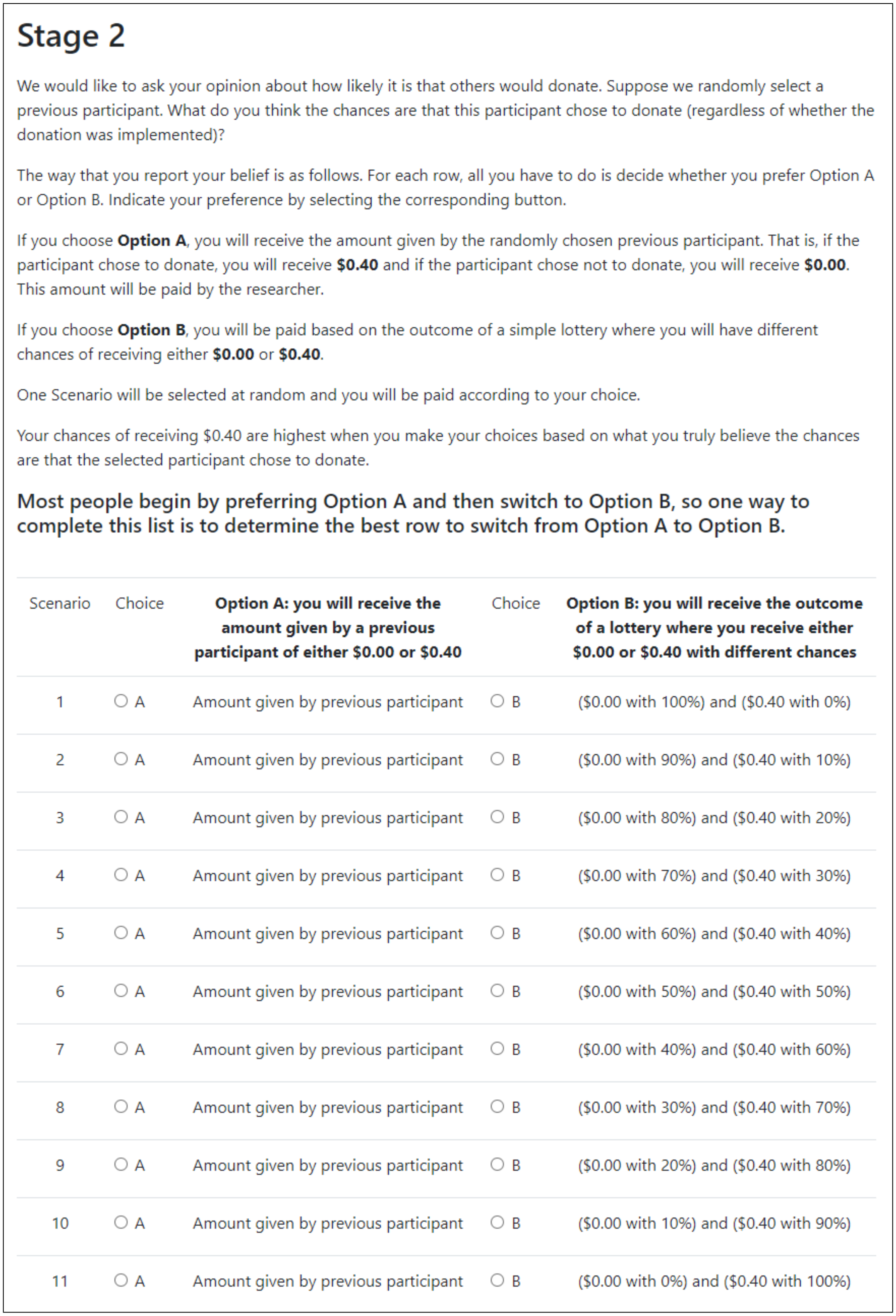
A previous group of 10 participants had the option of donating part of their earnings to a charity selected from the list below. A short description of each charity is also provided.

What do you believe was the charity most commonly chosen by the previous participants?

I believe the following charity was most commonly chosen:

Charity	Description
Against Malaria Foundation	Provides insecticide-treated nets to prevent malaria in sub-Saharan Africa
COVID Response Fund for WHO	Donations support WHO's work to track and understand the spread of the virus; to ensure frontline workers get essential supplies; and to accelerate research and development of a vaccine and treatments
Doctors without Borders	International humanitarian medical organisation with projects in conflict zones and in countries affected by endemic diseases
Feeding America	Non-profit organization that aims to feed people through food pantries, soup kitchens, shelters, and other community-based agencies
Johns Hopkins Centre for Health Security	Explores how new policy approaches, scientific advances, and technological innovations can stop pandemics, strengthen health security, and save lives
No Kid Hungry	Non-profit organization focused on alleviating childhood hunger in chaotic and stressful times
The Salvation Army	A Protestant christian church with charity shops, shelters for the homeless and offers disaster relief and humanitarian aid to developing countries
World Wildlife Fund	International organization working in the field of wilderness preservation, and the reduction of human impact on the environment

Inc-Karni-Exp



Stage 2

We would like to ask your opinion about how likely it is that others would donate. Suppose we randomly select a previous participant. What do you think the chances are that this participant chose to donate (regardless of whether the donation was implemented)?

The way that you report your belief is as follows. For each row, all you have to do is decide whether you prefer Option A or Option B. Indicate your preference by selecting the corresponding button.

If you choose **Option A**, you will receive the amount given by the randomly chosen previous participant. That is, if the participant chose to donate, you will receive **$0.40** and if the participant chose not to donate, you will receive **$0.00**. This amount will be paid by the researcher.

If you choose **Option B**, you will be paid based on the outcome of a simple lottery where you will have different chances of receiving either **$0.00** or **$0.40**.

One Scenario will be selected at random and you will be paid according to your choice.

Your chances of receiving $0.40 are highest when you make your choices based on what you truly believe the chances are that the selected participant chose to donate.

Most people begin by preferring Option A and then switch to Option B, so one way to complete this list is to determine the best row to switch from Option A to Option B.

Scenario	Choice	**Option A: you will receive the amount given by a previous participant of either $0.00 or $0.40**	Choice	**Option B: you will receive the outcome of a lottery where you receive either $0.00 or $0.40 with different chances**
1	○ A	Amount given by previous participant	○ B	($0.00 with 100%) and ($0.40 with 0%)
2	○ A	Amount given by previous participant	○ B	($0.00 with 90%) and ($0.40 with 10%)
3	○ A	Amount given by previous participant	○ B	($0.00 with 80%) and ($0.40 with 20%)
4	○ A	Amount given by previous participant	○ B	($0.00 with 70%) and ($0.40 with 30%)
5	○ A	Amount given by previous participant	○ B	($0.00 with 60%) and ($0.40 with 40%)
6	○ A	Amount given by previous participant	○ B	($0.00 with 50%) and ($0.40 with 50%)
7	○ A	Amount given by previous participant	○ B	($0.00 with 40%) and ($0.40 with 60%)
8	○ A	Amount given by previous participant	○ B	($0.00 with 30%) and ($0.40 with 70%)
9	○ A	Amount given by previous participant	○ B	($0.00 with 20%) and ($0.40 with 80%)
10	○ A	Amount given by previous participant	○ B	($0.00 with 10%) and ($0.40 with 90%)
11	○ A	Amount given by previous participant	○ B	($0.00 with 0%) and ($0.40 with 100%)

Chapter 2: Are three heads more biased than one? The role of communication in group belief updating[*]

Lata Gangadharan [†], Philip J. Grossman [‡], Nina Xue [§]

Abstract

Many economic decisions are made by teams, committees and boards. Individuals may come to these groups with different beliefs that must, ultimately, coalesce to a consensus belief. Yet, relatively little is known about how the beliefs that inform decision making in groups are formed and how these differ from individual beliefs. We conduct an experiment to examine the role of communication in belief updating. Overall, neither prior beliefs nor transfers differ between individuals and groups. Groups exhibit asymmetric updating but are not more biased than individuals. Based on text analyses, we identify risk preferences as an important topic in group communication and observe a self-serving bias in updating by more risk-averse groups – but not by risk-averse individuals. While the group environment does not necessarily lead to more motivated beliefs, communication can amplify individual preferences in a way that leads to more biased information processing by groups.

JEL Classification: C9, D9, H4
Keywords: belief updating, group decision making, self-serving bias, communication, experiment

[*] This work was supported by the Australian Research Council (DP170101167). We are grateful to Nick Feltovich, Roberto Weber, Sevgi Yuksel and Xiaojian Zhao and audiences at the University of Zurich Behavioral Lab Meeting and the 15th Australia and New Zealand Workshop on Experimental Economics for valuable feedback. This project received approval from the Monash University Human Research Ethics Committee (project number 23704).

[†] Monash University, Lata.Gangadharan@monash.edu

[‡] Monash University, Philip.Grossman@monash.edu

[§] Monash University, Nina.Xue@monash.edu

1 Introduction

A wide variety of important economic decisions are made by groups of decision makers, such as legislative and judicial committees, central banks and company boards. A majority of these decisions are made under uncertainty and require groups to aggregate the potentially different beliefs of individual members into a common group belief. Only once groups come to an agreement about the state of the world can they then decide on the best course of action. A natural question is whether important organizational decisions should be made by individual decision makers (e.g., the CEO) or by groups of decision makers (e.g., a board of directors). There is growing evidence that individual beliefs (e.g., Di Tella et al., 2015; Gino et al., 2016; Bicchieri et al., 2023) and belief updating (e.g., Eil and Rao, 2011; Coutts, 2019; Möbius et al., 2022), or the way in which new information is incorporated into beliefs, may be motivated by self-interest. However, it is not clear whether groups of individuals also exhibit such belief biases, or if the process of synthesizing individual beliefs affects the belief formation process. In this paper, we investigate the role of communication in group belief updating in the presence of a self-serving motive to bias beliefs.

Previous studies show that groups tend to make more self-interested decisions than individuals in social psychology (e.g., Insko and Schopler, 1987) and in a number of economic games measuring social preferences, including ultimatum games (e.g., Bornstein and Yaniv, 1998), dictator games (e.g., Luhan et al., 2009), centipede games (e.g., Bornstein et al., 2004), sender-receiver games (e.g., Behnk et al., 2022), and games measuring dishonesty (e.g., Kocher et al., 2018).[1] While this result holds for a majority of previous work, a number of papers also find no such effect, suggesting the effect may be context- and procedure-dependent.[2] In the context of the trust game, Cox (2002) finds that groups and individuals do not differ in the amount sent but groups return less, Kugler et al. (2007) show that groups send less than individuals but do not differ in the amount returned, while in Song (2008), group members are less trusting and trustworthy than individuals. One proposed explanation for why groups generally tend to be less cooperative than individuals, termed the "discontinuity effect", is "schema-based distrust" (Campbell, 1967; Insko and Schopler, 1987;

[1] In non-prosocial settings, group decisions have been shown to be closer to game-theoretic predictions due to a "wisdom of the crowd" effect (e.g., Cooper and Kagel, 2005; Charness et al., 2007; Fahr and Irlenbusch, 2011; Sutter et al., 2013; Cox and Stoddard, 2018). Kocher and Sutter (2005) find that groups learn faster than individuals in a beauty-contest game (guessing game), while Charness et al. (2007) show that groups are less likely than individuals to violate first-order stochastic dominance.

[2] Cason and Mui (1997) find that group decisions in the dictator game tend to be more altruistic than individual decisions when communication is face-to-face. In the sender-receiver game, Ambrus et al. (2015) find that group decisions do not differ from individual decisions and that median group members and those close to the median have significant impacts on the group decision. See also Charness and Sutter (2012) and Kocher et al. (2020) for reviews.

Schopler and Insko, 1992), the idea that the group decision-making environment activates an outgroup bias (e.g., Tajfel and Turner, 1979; Chen and Li, 2009) with groups believing that their opponents are less cooperative.[3] These beliefs, in turn, justify more self-interested behavior by groups. However, as highlighted in a review by Kugler et al. (2012), much of the literature has focused on giving behavior but has been vague regarding the beliefs of groups. To address this gap, we directly elicit the beliefs of individuals and groups and isolate the role of within-group communication on group beliefs. Understanding how groups form and update their beliefs is important for identifying any potential biases and improving decision making in group settings.[4]

We design a laboratory experiment using a simultaneous version (Costa-Gomes et al., 2014) of the trust game (Berg et al., 1995), in which beliefs have been shown to play an important role in transfer decisions (e.g., Guerra and Zizzo, 2004; Ashraf et al., 2006; Naef et al., 2008; Costa-Gomes et al., 2014). As a real-world example, many high-stakes negotiations in business and politics are conducted in groups and require mutual trust.[5] In the Group treatment (*Group*), participants make a joint transfer decision in groups of three, following communication via online chat. Before making this decision, group members agree on a common prior belief about their opponents' choices. We examine belief updating by eliciting three group posterior beliefs, each following a noisy binary signal (either low or high) about the likelihood that the group they are matched with transferred a high amount. In the Individual treatment (*Ind*), participants play the same trust game in groups of three, but transfer decisions are made individually (and without communication), on behalf of their group. We elicit the same prior and three posterior beliefs at the individual level. This is important in ensuring that the consequences of the decision are held constant across treatments, as decisions always affect a group, but we vary whether this decision is made by the group or by a group representative. We can therefore isolate the role of communication in belief formation and updating.

To check whether any observed differences in belief updating can indeed be attributed

[3] Another explanation is the "social support of shared self-interest", which proposes that group members offer mutual support for the pursuit of self interest that is independent of beliefs. Winquist and Larson (1998) find that making decisions as part of a group increases distrust of opponents, highlighting the interaction between the group decision making environment and beliefs.

[4] A related phenomenon that has been explored in the psychology literature is "groupthink" (Janis, 1972, 1983), the tendency for group members to conform to group values and a single way of thinking. While groupthink can manifest in a variety of ways, including group overconfidence (Bénabou, 2013) and the censoring of alternative views, we examine a specific type of groupthink that involves biased beliefs that help to justify self-serving behavior.

[5] For instance, a successful ceasefire agreement requires trust from both sides, despite the temptation to reduce one's own vulnerabilities and act selfishly. One possible justification for self-interested behavior is the belief that the other side is not trusting or trustworthy.

to a self-serving bias, we conduct the No-Transfer treatments (*Group-NT* and *Ind-NT*), in which we elicit the beliefs of disinterested third parties who do not participate in the trust game and thus have no personal stake in the outcome.[6] The No-Transfer treatment allows us to "switch off" the self-serving motive as these participants have no transfer decision to make and hence have no personal incentive to bias their beliefs. This also helps to control for any differences in information processing between individuals and groups that are not driven by self-interest.

Our analysis is guided by a simple conceptual framework in which agents are motivated to minimize the cognitive dissonance between a (self-interested) desire to both maximize monetary payoffs and to perceive their own actions as fair (Festinger, 1957; Akerlof and Dickens, 1982; Rabin, 1994; Konow, 2000). Reducing this dissonance can be achieved either by giving more, or by holding a belief that the opponent made a low transfer, thus justifying a low transfer in return. We directly test whether communication plays a key role in changing group beliefs and conjecture that the psychological cost of distorting beliefs is lower for groups than for individuals. This implies that groups are more likely to both hold self-serving beliefs and make lower transfers, as cognitive dissonance is alleviated through the beliefs channel, placing less pressure on transfers.

Experimental methods are particularly useful in enabling us to control for many differences that typically exist between the individual and group decision-making environments and isolate the role of communication. Second, experiments allow us to manipulate whether an agent has a vested interest in the outcome, about which we elicit beliefs. Through this, we can identify whether deviations from Bayesian updating are truly due to motivated biases, or due to other factors, such as cognitive limitations. A third advantage of experiments is that we can obtain a direct measure of beliefs (that is not typically found in empirical data) whilst controlling for the informativeness of signals given to participants.

Consistent with previous work, we find that beliefs positively predict transfer decisions in the trust game. However, we find no evidence that group prior beliefs or group transfers differ from that of individuals. Using the information processing framework introduced by Grether (1980) and developed by Möbius et al. (2022), we observe asymmetric updating in groups (but not in individuals), as group beliefs are more responsive to low signals (that suggest the other group did not transfer a high amount) than high signals. However, overall, we are not more likely to observe a self-serving bias in updating by groups than by individuals.

To explore the content of group communication (Cooper and Kagel, 2005; Brandts et al., 2019; Gentzkow et al., 2019), we conduct a text analysis whereby we organize the most

[6] See, for example, Babcock et al. (1995) and Konow (2000) for similar designs featuring a neutral third-party observer.

frequently-observed relevant keywords into topics. We also support this analysis using Latent Dirichlet Allocation (LDA), an unsupervised machine learning algorithm which organizes chat messages into an optimal number of natural groups based on how likely certain keywords appear together in the chat (Blei et al., 2003).[7] Both analyses identify risk as an important factor in groups' decisions. We subsequently investigate whether risk preferences play a role in belief updating and find evidence of a self-serving bias by more risk-averse groups (when comparing beliefs across the Transfer and No-Transfer treatments), but not by more risk-tolerant groups. This bias is consistent with the desire to reduce one's own risk exposure by believing the other group to be distrusting. In contrast, we do not observe any self-serving biases by more risk-averse individuals. The group setting, as well as the composition of a group, therefore, both matter for the way in which groups respond to new information and aggregate their beliefs.

One implication of our findings is that the group decision-making environment *per se* does not lead to more biased belief updating. In general, group decisions are not significantly different from those of individual decision-makers. Given the other documented benefits of decision making in teams (e.g., Bainbridge, 2002; Glassop, 2002), and the tendency for groups to make more rational decisions and learn faster than individuals (e.g., Cooper and Kagel, 2005; Maciejovsky et al., 2013), we offer no categorical evidence against organizations delegating important decisions to teams. However, our findings do highlight the ability of the group environment to amplify individual preferences (such as risk preferences) in a way that could lead to more biased beliefs. Our results therefore caution that a lack of diversity in the institutional environment may create conditions that facilitate more biased decision making and offer a case for a greater variety of perspectives in teams.

Our research complements recent evidence that individual belief updating can become more biased following a social exchange of beliefs. Oprea and Yuksel (2020) allow participants to observe their partner's real-time belief adjustments and find that beliefs about own ability become more upwardly biased, or overconfident, over time. Similarly, Kogan et al. (2021) examine beliefs about group outcomes and show that individual beliefs tend to become more biased following an exchange of information via a market mechanism. Both papers focus on beliefs in the domain of ability and do not allow for free-form communication within groups. Mengel (2021) studies the impact of communication within a hiring committee and finds that evaluations tend to be more biased against women following group deliberation. However, Mengel (2021) does not explicitly examine how groups respond to new information, or allow for the possibility of self-serving biases, both of which are a key focus of this study.

[7] See, Penczynski (2019) for the use of machine learning to analyze communication data and Hanaki and Ozkes (2023) and Andres et al. (2022) for LDA analyses in economics.

2 Experimental design

We conduct a laboratory experiment in which participants play a simultaneous trust game (Costa-Gomes et al., 2014) in groups. Half of the participants are assigned the role of trustor and are randomly sorted into groups of three ("Group A"). The other half are assigned the role of trustee and are also randomly sorted into groups of three ("Group B"). Both groups start with the same endowment of 100 ECU.[8] Trustors send a share of their endowment (0 - 100%) to the trustees they are randomly matched with, and this amount is tripled by the experimenter before it is added to the trustees' group account. Before knowing the total ECUs in their group account, trustees decide on a share of their account balance (0 - 100%) to send back to the trustors they are matched with. The final account balance is shared equally among group members. There is robust evidence on the importance of beliefs about opponents in trust games (e.g., Guerra and Zizzo, 2004; Ashraf et al., 2006; Naef et al., 2008; Costa-Gomes et al., 2014); we chose the simultaneous version of the game because beliefs play an even more important role, since trustees are not informed of the trustor's exact transfer before making their own transfer decision.

We employ a 2 x 2 between-subjects design, with participants assigned to one of the following treatments: Individual (*Ind*), Group (*Group*), Individual No-Transfer (*Ind-NT*) and Group No-Transfer (*Group-NT*), see Appendix A for screenshots of the instructions. The only treatment difference is that participants in *Group* can communicate with their group members before making a joint group decision, while in *Ind*, individual group members make the same decision on *behalf* of their group. Previous work investigating the discontinuity effect have compared the effect of playing against a group versus playing against an individual and playing as part of a group versus playing as an individual (Winquist and Larson, 1998; Kugler et al., 2007). One concern with such asymmetric designs, however, is that it is difficult to control for other confounds when comparing across individuals and groups, e.g., efficiency concerns or altruistic concerns for group members. Our design holds the consequences of the decision constant and only varies whether the decision is made by a single decision maker or a group. Another advantage of our design is that we can isolate the role of communication in the formation of beliefs and the decision-making process. We directly elicit incentivized beliefs from groups after allowing for group discussion and compare these beliefs against those of individuals representing their groups.

In the No-Transfer treatments, we elicit beliefs from third-party observers who do not participate in the game, thereby reducing the self-serving motive to distort beliefs (see Appendix B for a summary of key treatment differences). We first explain *Group*, followed by

[8] This reduces the role of inequality aversion on transfer decisions in the game.

how *Ind* differs. We then describe how the No-Transfer treatments differ from the Transfer treatments.

2.1 Group treatment

2.1.1 Transfers

Participants in *Group* can communicate with their group members via online chat and make a joint transfer decision. In order to encourage groups to come to an agreement, and following previous work on group decision making (e.g., Luhan et al., 2009; Cox and Stoddard, 2018), we enforce a unanimity rule whereby each group member must enter the same number within the allocated time for the group's decision to be valid. Groups have 4 minutes to enter a valid group transfer. If no valid group transfer has been entered, then the computer implements a default transfer of 70%. We chose a high default to increase the stakes of disagreement.[9]

2.1.2 Beliefs

Since beliefs are the main focus of this paper, we first elicit subjects' beliefs about the decisions of their counterparts, before eliciting transfer decisions.[10] First, for each member of Group A, we elicit individual prior beliefs about the proportion of Group B's in the following session (one of which they will be matched with) that will transfer a quarter (25%) or more of their group's account back to Group A.[11] In other words, subjects are asked for the likelihood that they will be matched with a trustee group that transfers a relatively higher amount. The two main advantages of asking beliefs in this way are that the mechanism is easy to understand and beliefs can be incentivized based on the choices of other participants.[12] Similarly, for each member of Group B, we elicit individual prior beliefs about the proportion of Group A's that transfer a quarter or more of their group's endowment. Second, we elicit the prior beliefs of groups using the same belief question as above, only this time subjects can communicate with their group members to arrive at a single group belief. We require each group member to enter the same value for the group belief to be valid.[13]

Finally, we elicit three group posterior beliefs from Group B's, each following a noisy

[9] In the experiment, all groups agreed on a transfer and this default was never applied.

[10] Gangadharan et al. (2022) find that the order of elicitation does not affect beliefs nor choices in an individual donation task.

[11] We chose this value based on a pilot with four Group A's and Group B's (N=24), in which the median transfer was approximately 25%.

[12] See Schlag and Tremewan (2021) and Gangadharan et al. (2022) for similar methods and Charness et al. (2021) for a discussion of complex versus simple belief elicitation methods.

[13] If no valid belief has been entered within four minutes, the computer implements a belief of 0%. Four Group A's and six Group B's failed to reach an agreement for the group prior belief.

signal about the amount transferred by trustors in the previous session. Note that we always conducted two sessions in succession – the first with Group A's, and the second with the corresponding Group B's. This ordering was crucial in providing noisy but truthful signals to Group B's about the actual decisions of Group A's. We chose to focus on Group B's beliefs as there are fewer behavioral confounds, e.g., Group A's decision may also be driven by efficiency concerns. The three independent signals are delivered via a "Magic-8-Ball" which displays one of two messages: either that the proportion of Group A's that transferred a quarter or more was "Greater or equal to 33%" (i.e., a high signal), or "Less than 33%" (i.e., a low signal).[14] However, subjects are informed that the Magic-8-Ball can be faulty at times and will reveal the wrong message with a 1 in 3 chance, i.e., each observed signal has a 2/3 chance of being correct. The signals are informative but not perfectly so, in order to inject uncertainty into the belief updating process. Following each message from the Magic-8-Ball, group members can use the online chat to decide on a common group posterior belief. Given that we can verify actual transfer decisions from the corresponding session, all beliefs are incentivized such that participants receive an additional $10 if their guess is within ±5% of the actual proportion.

2.2 Individual treatment

2.2.1 Transfers

In order to keep the incentive structure comparable in our study between *Group* and *Ind*, participants in *Ind* are similarly matched in groups of three, but make transfer decisions on *behalf* of their group. Each group member makes an individual decision about how much of their group's account to transfer to the group they are matched with, without consulting their group members. Similar to *Group*, participants have 4 minutes to enter a value and if no transfer has been entered, then the computer implements a default of 70%. At the end of the study, one group representative is chosen at random and their transfer decision is implemented for the group. Thus, we can compare decisions that are made following group discussions, against decisions that are made by an individual decision maker, whilst still controlling for efficiency and welfare concerns.

[14] A Magic-8-Ball is a children's toy that is designed to be used as a fortune-telling device. The user can ask a "yes/no" question and upon turning the ball, a randomly chosen message from a finite set of possibilities appears in the window. We chose the cutoff of 33% based on a pilot (N=21) in which participants were only asked for their individual beliefs.

2.2.2 Beliefs

Regarding beliefs, the only departure from *Group* is that in *Ind*, participants are asked to privately report their individual beliefs. For Group A members, we elicit their individual prior beliefs about the proportion of Group B's that transfer a quarter or more of their group's account. Similar to *Group*, participants in *Ind* also have 4 minutes to enter a valid belief, otherwise the computer implements a default belief of 0%. For Group B members, we elicit their individual prior beliefs about the proportion of Group A's that transfer a quarter or more of their endowment, and three individual posterior beliefs following three noisy signals from the Magic-8-ball with the same accuracy rate of 2/3. Similar to the group treatment, beliefs were also incentivized for accuracy.

2.3 No-Transfer treatments

To test whether beliefs are different in the absence of a self-serving motive, we compare the No-Transfer treatments against *Group* and *Ind*. In *Group-NT* and *Ind-NT*, participants are explicitly informed at the beginning of the experiment that they will not make any transfer decisions, and are only asked for their beliefs about the decisions made by previous participants. In *Group-NT*, we elicit individual prior beliefs, group prior beliefs and three group posterior beliefs about the proportion of Group B's from a previous session that transferred a quarter or more of their group's account. In *Ind-NT*, we elicit individual prior beliefs and three individual posterior beliefs, all incentivized for accuracy. Since our main focus is on belief updating, subjects are asked the same belief questions that Group B members were asked, i.e., prior and posterior beliefs regarding the choices of Group A's from a previous session.

2.4 Risk preferences and post-experiment survey

An important element of trust is risk, as trustworthiness requires individuals to go against their own self-interest. Previous studies have shown that individual risk attitudes are an important driver of trusting behavior (e.g., Karlan, 2005; Schechter, 2007; Fehr, 2009). Trust may also entail an additional element of social risk, i.e., the possibility that the trust placed in another individual is betrayed (e.g., Bohnet and Zeckhauser, 2004; Houser et al., 2010). In the context of the simultaneous trust game, both Group A and B are vulnerable to the risk of their trust not being reciprocated as either group can choose to transfer 0% or a very low amount. We elicit individual risk attitudes using the Eckel and Grossman (2002) measure in the post-experiment survey to investigate the role of risk preferences in belief and transfer decisions. Participants also complete a demographic survey on gender, age, ethnicity, religiosity, income and political orientation, and some open-ended questions about their decision-making process during the experiment.

2.5 Procedures

The experiment was programmed using oTree (Chen et al., 2016). Our experimental design and research questions were pre-registered on AsPredicted.org (pre-registration #74705). Sessions were conducted at the Monash Laboratory for Experimental Economics at Monash University, using Sona to recruit subjects, and the Experimental Economics Laboratory at the University of Melbourne, using ORSEE (Greiner, 2015) for recruitment, between August - October 2021.[15] Due to stay-at-home orders in Melbourne, we conducted online sessions of 18 subjects per session via Zoom with conditions similar to a laboratory environment.[16]

In all treatments, participants start with a team-building task, in which they solve six problems based on the Remote Associations Test (RAT), a measure of creativity by Mednick (1962). Group members can use the online chat function to solve the problems together. This task is used to firstly familiarize participants with the online chat, and secondly, to ensure that all groups (including those in *Ind*) have the same opportunity for team bonding (see e.g., Eckel and Grossman, 2005). This reduces the likelihood that participants in *Group* care more about their group members' payoffs (and hence may keep more for their own group) after the opportunity to bond via the online chat. Subjects received a show-up fee of 10 AUD plus payment for either their outcome in the trust game, or for one of their beliefs, and this was determined at random.

We report results from a total of 360 participants (with N=240 in the Transfer treatment and N=120 in the No-Transfer treatment).[17] To ensure subjects understood the instructions, we included comprehension questions on the trust game, the signals by the Magic-8-Ball and the group decision-making mechanism (in the group treatments). Participants earned $16.86 on average and the experiment lasted less than one hour.

[15] Our results do not differ across the Monash University and the University of Melbourne subject pools.

[16] For example, sessions were anonymized, instructions were read out loud by the experimenter, subjects could ask private questions to the experimenter and had to correctly answer comprehension questions before proceeding to the decision screens.

[17] As specified in our pre-registration, we collected 30 independent observations for each player type per treatment. In the No-Transfer treatments, we had no need for Player A's since no transfer decision was made.

3 Conceptual framework

Our goal in this section is twofold. First, we present a simple conceptual framework outlining the self-serving motive of agents to both maximize monetary payoffs and perceive their own action as fair. Second, we clarify why beliefs may differ across groups and individuals in the presence of this self-serving motive to bias beliefs.

As mentioned in Section 2, we focus our analysis on Group B's beliefs, whose utility consists of three terms.[18] The first is monetary utility, $\pi(x_B)$, which is assumed to be twice continuously differentiable in the proportion of B's endowment that is transferred (x_B), with $\pi' < 0$ and $\pi'' < 0$. Second, agents are motivated to reduce the cognitive dissonance (Festinger, 1957; Akerlof and Dickens, 1982; Rabin, 1994) between a desire to maximize payoffs and a desire to believe their transfer is fair or socially appropriate. Based on a model outlined by Konow (2000), let $n(\hat{x}_A) \in [0, 1]$ denote the socially appropriate proportion that B should send, or the social norm (Cialdini et al., 1990; Bicchieri, 2005), which depends on their beliefs about A's transfer $(\hat{x}_A)$.[19] We assume that $n(\hat{x}_A)$ is twice continuously differentiable in $\hat{x}_A$, with $n' > 0$ and $n'' < 0$. The more B believes was transferred by A, the higher the socially appropriate proportion B should transfer back to A. When B believes A transferred zero, then we assume it is socially appropriate to also transfer zero, $n(0) = 0$. Let $w = n(\hat{x}_A) - x_B$ represent the difference between the socially appropriate transfer and B's actual transfer. As the difference between the norm and B's transfer increases, B experiences more cognitive dissonance, $f(w)$, or disutility from violating the social norm. Cognitive dissonance is assumed to be twice continuously differentiable in w, with $f(\cdot)$ increasing in w and is a strictly convex function of w, increasing at an increasing rate as the difference between $n(\hat{x}_A)$ and x_B becomes larger. Thus far, B would maximize their utility by transferring zero $(x_B = 0)$, supported by their belief about A's transfer $(\hat{x}_A = 0)$.

In reality, however, there are generally limits to what one can reasonably believe. The third term is thus a psychological cost of distorting beliefs. Let $\bar{x}_A$ represent the belief held by a disinterested third party (with no stake in the transfer decision) about A's transfer. The psychological cost, $c(y, \beta)$, depends on the difference between what a reasonable outsider would believe and B's beliefs, $y = \bar{x}_A - \hat{x}_A$, and is assumed to be continuously differentiable in β, $\beta \neq 1$, twice continuously differentiable in y, $\beta \neq 1$, is increasing in y and is a strictly convex function of y. The larger this difference, the greater the belief distortion and the higher the psychological costs of convincing oneself that this belief is correct (Kunda, 1990).

[18] The analysis for Group A is symmetric.

[19] For example, if the norm is that B should transfer some amount that ensures A is not made worse off than their original endowment, then $n(\hat{x}_A) = \frac{\hat{x}_A}{3\hat{x}_A+1}$.

For instance, it may take more cognitive effort to selectively recall arguments that support a biased belief (e.g., Chew et al., 2020; Zimmermann, 2020). The psychological cost is also increasing in β, which is a parameter that indicates how costly it is to distort beliefs and may vary across individuals and contexts. We assume $\beta > 0$, i.e., the cost of distorting beliefs is not zero.

B chooses the levels of two variables, how much to transfer (x_B), and their belief about A's transfer ($\hat{x}_A$) and solves the following problem:

$$\begin{gathered}\max_{x_B,\hat{x}_A} u(x_B,\hat{x}_A,\beta) \equiv \pi(x_B) - f(n(\hat{x}_A) - x_B) - c(\bar{x}_A - \hat{x}_A,\beta) \\ \text{subject to } 0 \leq x_B \leq 1, 0 \leq \hat{x}_A \leq 1\end{gathered} \tag{1}$$

At the optimum, marginal dissonance equals marginal belief distortion costs, and equals marginal monetary utility.

Proposition 1: $\frac{\partial x_B^*}{\partial \beta} \geq 0$ and $\frac{\partial \hat{x}_A^*}{\partial \beta} \geq 0$

Proposition 1 means that when belief distortion is more costly, B's beliefs are less biased, which in turn increases B's transfer. We conjecture that the psychological costs of belief distortion are lower in groups following communication, $\beta^{Group} < \beta^{Ind}$. Allowing group members to communicate with one another may help groups to come up with more arguments for why A's transfer might be low, or self-serving arguments that support a low belief may receive more validation from other group members. These lower psychological costs imply that beliefs are lower for groups than for individuals ($\hat{x}_A^{Group} < \hat{x}_A^{Ind}$). This is consistent with survey evidence by Winquist and Larson (1998) that the group decision-making environment increases distrust of opponents, and Kugler et al. (2007), who show that group members have lower expectations about returns than individuals.[20] As a consequence, this means that groups are better able to reduce cognitive dissonance through belief distortion, which places less pressure on them to make a high transfer ($x_B^{Group} < x_B^{Ind}$). This prediction is in line with previous work documenting the tendency for groups to make more self-interested decisions than individuals (e.g., Bornstein and Yaniv, 1998; Song, 2008; Luhan et al., 2009).

In addition to examining prior beliefs, the focus of this paper is on how groups and individuals process new information, which we analyze within a Bayesian-updating framework. In the context of our study, we focus on the *relative* responsiveness to low and high signals (i.e., asymmetry) as an indication of biased belief updating. We conjecture that within-group

[20] Beliefs in Kugler et al. (2007) were elicited from each individual group member (without within-group discussion) and groups were not required to agree on a single group belief.

communication may help groups to discount high signals and over-weigh low signals due to belief distortion being less costly in groups.

Two broad explanations have been proposed in the literature (e.g., Benjamin, 2019; Coutts, 2019; Möbius et al., 2022) for why posterior beliefs may deviate from the Bayesian benchmark: (1) motivated or self-serving biases (which Benjamin (2019) terms "preference-biased updating"), and (2) cognitive limitations in belief updating. Given that Bayesian updating is not straightforward, agents may simply lack the cognitive capacity to apply the necessary calculations. These cognitive limitations may be unrelated to a desire to hold a self-serving belief. We conduct the No-Transfer treatments in order to remove any self-serving motives due to participants' personal stakes in the decision and outcome. The beliefs of these third parties are simply $\bar{x}_A$. We subsequently compare belief updating in the Transfer and No-Transfer treatments to check if any observed differences between groups and individuals can be attributed to self-interest. Specifically, if we do not observe a similar difference between *Group-NT* and *Ind-NT*, this would suggest that deviations from Bayesian updating in the Transfer treatments are driven by a self-serving bias. Conversely, if we observe no difference between how second and third parties process information, this would suggest that any differences between individuals and groups are more likely to be cognitive in nature.

4 Results

4.1 Prior beliefs

In this section, we first investigate whether prior beliefs change after allowing for communication in groups. In order to isolate the role of communication in group prior beliefs, we first examine whether individuals expect group representatives to differ from groups in the amount transferred in the trust game, i.e., we compare individual prior beliefs across *Group* and *Ind*.[21] According to a two-tailed Mann-Whitney test, individual priors do not differ significantly between individuals and groups for Group A (*Ind*: 39.33 vs. *Group*: 38.48, $p = 0.76$) and Group B (34.87 vs. 39.63, $p = 0.38$).[22] This is consistent with the regression analysis presented in Table 1 for Group A ($p = 0.97$, column 2), Group B ($p = 0.32$, column 4) and the two groups pooled ($p = 0.58$, column 6).

We next focus on participants in *Group* to compare prior beliefs before (i.e., individual priors) and after communication (i.e., group priors). Individual and group prior beliefs do

[21] Given that the trust game is not symmetric and Group A and Group B make different decisions, we first analyze these beliefs separately before pooling the beliefs together.

[22] Unless otherwise specified, we use a two-tailed Mann-Whitney test to compare average beliefs and transfers.

Table 1: Individual prior beliefs

	Group A		Group B		Pooled	
	(1)	(2)	(3)	(4)	(5)	(6)
Group	-0.86	-0.21	4.77	5.97	1.96	2.16
	(5.46)	(5.71)	(5.37)	(6.03)	(3.82)	(3.95)
Constant	39.33^{***}	141.14^{***}	34.87^{***}	16.18	37.10^{***}	114.52^{***}
	(4.73)	(29.92)	(4.65)	(23.52)	(3.30)	(28.37)
Controls	No	Yes	No	Yes	No	Yes
R^2	0.00	0.35	0.01	0.24	0.00	0.16
Adj. R^2	-0.01	0.14	-0.00	0.01	-0.00	0.04
Num. obs.	120	120	120	120	240	240

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the individual's prior belief about the proportion of groups sending a quarter of more of their group account. The baseline Treatment is *Ind*. The control variables are: gender, age, ethnicity, religiosity, income, political ideology and risk preferences. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

not differ significantly for Group A (individual: 37.74 vs. group: 34.85, $p = 0.65$) or for Group B (41.04 vs. 42.75, $p = 0.39$). The regression analysis in Table 2 confirms this result for Group A ($p = 0.45$, column 1), Group B ($p = 0.66$, column 2) and when we pool the beliefs of Group A and Group B ($p = 0.80$, column 3): Communication, therefore, does not appear to shift average prior beliefs. Reported confidence levels in prior beliefs do not differ significantly between individuals and groups for Group A's (individual: 6.09 vs. group: 6.16, $p = 0.91$) or for Group B's (5.57 vs. 5.92, $p = 0.27$).[23]

Table 2: Individual and group prior beliefs in *Group*

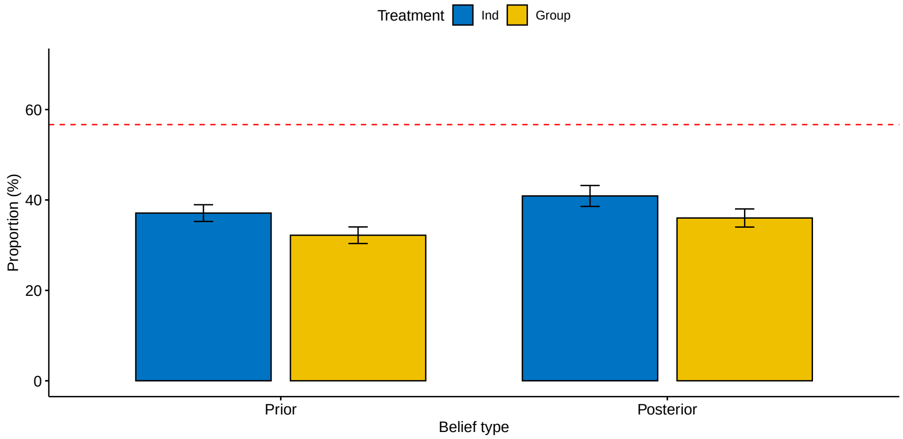
	Group A	Group B	Pooled
	(1)	(2)	(3)
Group prior	-2.90	1.71	-0.69
	(3.86)	(3.93)	(2.77)
Constant	37.74^{***}	41.04^{***}	39.33^{***}
	(2.73)	(2.78)	(1.96)
R^2	0.00	0.00	0.00
Adj. R^2	-0.00	-0.01	-0.00
Num. obs.	156	144	300

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the prior belief about the proportion of groups sending a quarter of more of their group account. The baseline belief is the individual prior. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

[23] While confidence in prior beliefs does not differ across treatments, confidence in beliefs is positively correlated with prior and posterior beliefs for both individuals and groups, see Appendix C.

From our sample, the true proportion of individuals in Group A transferring a quarter or more of their endowment was 56.67% and for groups this was also 56.67%, see Figure 1. Both individual prior beliefs (Wilcoxon signed-rank test, $p < 0.01$) and group prior beliefs (Wilcoxon signed-rank test, $p < 0.01$) significantly underestimate the true proportion.

Figure 1: Prior and posterior beliefs by Group B's



Notes: Error bars represent standard errors. The red dotted line represents the true proportion of Group A's sending a quarter or more of their endowment (56.67%).

Result 1: Following communication, group prior beliefs are not significantly different from individual prior beliefs.

4.2 Belief updating

In this section, we investigate how individuals and groups incorporate new information into their beliefs and the role of within-group communication in the updating process. We first describe the framework we use to compare observed updating against the Bayesian benchmark. Then, we present our results on posterior beliefs by groups and individuals in the Transfer treatment, followed by a comparison of updating by second (in the Transfer treatments) and third parties (in the No-Transfer treatments).

In general, agents respond to signals in the expected direction: Posterior beliefs following a high signal are greater than those following a low signal, for both individuals (high: 43.64 vs. low: 33.35, $p < 0.01$) and groups (40.56 vs. 32.59, $p < 0.01$).[24] Similar to prior beliefs, both individuals and groups underestimate the true proportion of Group A's that send more than a quarter of their endowment in their posterior beliefs (Wilcoxon signed-rank test,

[24] See Appendix D for a more detailed analysis.

both $p < 0.01$, see Figure 1). Confidence in posterior beliefs do not vary significantly across treatments (*Ind*: 6.16 vs. *Group*: 5.93, $p = 0.64$). This result holds for both low (5.96 vs. 5.93, $p = 0.96$) and high signals (6.23 vs. 5.92, $p = 0.15$).

4.2.1 Information processing framework

We follow the framework introduced by Grether (1980) and developed by Möbius et al. (2022) to estimate agents' responsiveness to the signals, and compare this against Bayesian updating.[25] Given binary signals about the proportion of previous groups that are either high (H), i.e., "Greater or equal to 33%", or low (L), i.e., "Less than 33%", Bayes' rule can be written in the following form:

$$logit(\hat{\mu_t}) = logit(\hat{\mu}_{t-1}) + \mathbf{1}(s_t = H)ln(\lambda_H) + \mathbf{1}(s_t = L)ln(\lambda_L) \tag{2}$$

where $\hat{\mu_t}$ is the posterior belief at time t, $\mathbf{1}(.)$ is an indicator function for the type of signal observed, and λ_{s_t} is the likelihood ratio of observing the signal $s_t \in \{H, L\}$. In our study, $\lambda_H = 2$ and $\lambda_L = 1/2$, based on an accuracy rate of 2/3. We estimate the following regression, which nests the Bayesian benchmark from (2) as a special case:

$$logit(\hat{\mu_{it}}) = \delta^{prior} logit(\hat{\mu}_{i,t-1}) + \beta^{high}\mathbf{1}(s_{it} = H)ln(\lambda_H) + \beta^{low}\mathbf{1}(s_{it} = L)ln(\lambda_L) + \epsilon_{it} \tag{3}$$

where δ^{prior} denotes the weight placed on the prior belief, β^{high} and β^{low} capture responsiveness to high and low signals, respectively, and ϵ_{it} captures non-systemic errors. Bayesian updating would predict that $\delta^{prior} = \beta^{high} = \beta^{low} = 1$, i.e., a Bayesian agent places equal weights on their prior belief, a high signal, and a low signal, after accounting for signal accuracy. In our analysis, we can examine whether groups and individuals place appropriate weights on their prior beliefs, or if they exhibit base-rate neglect ($\delta^{prior} < 1$), or a confirmation bias ($\delta^{prior} > 1$). Our main focus, however, is on the comparison of *relative* (as opposed to absolute) sensitivity to high and low signals, to determine whether groups and individuals respond to new information in a symmetric way. In particular, we test whether agents exhibit asymmetric updating ($\beta^{high} \neq \beta^{low}$), which would indicate a systematic bias in information processing (e.g., subjects may discount high signals more than low signals if they are motivated to believe that the other group did not send very much).

[25] See e.g., Coutts (2019), Oprea and Yuksel (2020), Erkal et al. (2021) and Castagnetti and Schmacker (2022) for similar analyses using this framework.

4.2.2 Belief updating in the Transfer treatments

We first examine belief updating in the Transfer treatment separately for individuals and groups in columns 1 and 2 of Table 3. Consistent with previous work on belief updating (e.g., Coutts, 2019; Erkal et al., 2021; Möbius et al., 2022), we find evidence of base-rate neglect ($\delta^{prior} < 1$), i.e., participants place less weight on their prior beliefs compared to a Bayesian agent, in both *Ind* ($p < 0.01$) and *Group* ($p < 0.01$).[26]

Turning to the comparison between high and low signals, we find that individuals tend to under-weigh both high ($p < 0.01$) and low signals ($p = 0.06$). Groups, on the other hand, are less responsive to high signals ($p < 0.01$) but do not differ from Bayesians in their response to low signals ($p = 0.62$). We observe asymmetric updating in *Group*, as groups tend to place more weight on low signals relative to high signals ($p < 0.01$), but this asymmetry is not observed in individuals ($p = 0.58$). Next, we pool data from *Ind* and *Group* to test whether updating is *more* asymmetric in groups than in individuals and find some evidence of more asymmetry in groups ($p = 0.09$, column 3).

Result 2: Following communication, groups exhibit asymmetric updating, placing more weight on low signals and less weight on high signals. We find no evidence of asymmetric updating in individuals.

4.2.3 Belief updating in the No-Transfer treatments

To determine whether the asymmetric updating observed in groups (but not by individuals) is motivated by self-interest, we examine the posterior beliefs of third parties. We conjecture that in the absence of a self-serving motive, participants in the No-Transfer treatments have a weaker incentive to update their beliefs in a biased way. We therefore examine whether belief updating in groups is asymmetric due to the self-serving motive, or simply because groups and individuals process information differently.

Columns 1 and 2 of Table 4 present separate regression analyses for *Ind-NT* and *Group-NT*, showing that similar to the Transfer treatments, third-party individuals and groups under-weigh prior beliefs (both $p < 0.01$) and high signals (both $p < 0.01$). However, neither individuals ($p = 0.77$) nor groups ($p = 0.41$) differ from Bayesian agents in their response to low signals. We observe asymmetric updating by third-party groups ($p = 0.02$) and also find some evidence of asymmetry in third-party individuals ($p = 0.07$). However, unlike

[26] In Appendix E, we exclude beliefs in which updating occurs in the wrong direction, which make up 7% of beliefs by groups and individuals. Appendix F presents the same analysis after using a proxy for group beliefs where no agreement was reached.

Table 3: Belief updating in *Ind* and *Group*

	Ind (1)	*Group* (2)	Pooled (3)
δ^{prior}	0.54***	0.25***	0.54***
	(0.15)	(0.09)	(0.16)
β^{high}	0.21	−0.34*	0.21
	(0.27)	(0.21)	(0.27)
β^{low}	0.42*	0.94***	0.42*
	(0.25)	(0.16)	(0.25)
δ^{prior} x *Group*			−0.31*
			(0.18)
β^{high} x *Group*			−0.53
			(0.33)
β^{low} x *Group*			0.45
			(0.29)
$H_0 : \delta^{prior} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{high} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p = 0.02$	$p = 0.71$	$p = 0.02$
$H_0 : \beta^{high} = \beta^{low}$	$p = 0.67$	$p < 0.01$	$p = 0.67$
$H_0 : \beta^{high}$ x *Group* = β^{low} x *Group*	-	-	$p = 0.09$
R^2	0.45	0.49	0.47
Adj. R^2	0.45	0.49	0.46
Num. obs.	270	249	519

Notes: Ordinary least squares regression with standard errors clustered at the individual level (in *Ind*) and at the group level (in *Group*) in parentheses. The dependent variable is the belief about the proportion of groups that send a quarter or more of their account. We use individual prior beliefs for *Ind* and group prior beliefs for *Group*. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

with second parties, belief updating by groups is not significantly more asymmetric than by individuals ($p = 0.96$, column 3).

Table 5 pools observations across the Transfer and No-Transfer treatments. Overall, we do not find any evidence of a self-serving bias in individuals or groups in the Transfer treatments, compared to the third-party benchmark. Individuals in *Ind* and *Ind-NT* place similar weights on their prior beliefs ($p = 0.20$, column 1), high signals ($p = 0.27$), and low signals ($p = 0.40$). Similarly for groups, we do not find a significant difference in the weights assigned to group prior beliefs ($p = 0.62$, column 2), high signals ($p = 0.71$), or low signals ($p = 0.76$).

Result 3: Overall, we do not find evidence of a self-serving bias in belief updating for individuals or for groups.

Table 4: Belief updating in *Ind-NT* and *Group-NT*

	Ind-NT (1)	*Group-NT* (2)	Pooled (3)
δ^{prior}	0.29^{**}	0.31^{**}	0.29^{**}
	(0.12)	(0.14)	(0.12)
β^{high}	-0.17	-0.21	-0.17
	(0.21)	(0.21)	(0.22)
β^{low}	0.87^{*}	0.78^{***}	0.87^{*}
	(0.46)	(0.26)	(0.46)
δ^{prior} x *Group*			0.02
			(0.19)
β^{high} x *Group*			-0.04
			(0.30)
β^{low} x *Group*			-0.08
			(0.53)
$H_0 : \delta^{prior} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{high} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p = 0.77$	$p = 0.41$	$p = 0.77$
$H_0 : \beta^{high} = \beta^{low}$	$p = 0.07$	$p = 0.02$	$p = 0.07$
$H_0 : \beta^{high}$ x *Group* = β^{low} x *Group*	-	-	$p = 0.96$
R^2	0.19	0.35	0.25
Adj. R^2	0.18	0.34	0.24
Num. obs.	270	234	504

Notes: Ordinary least squares regression with standard errors clustered at the individual level (in *Ind-NT*) and at the group level (in *Group-NT*) in parentheses. The dependent variable is the belief about the proportion of groups that send a quarter or more of their account. We use individual prior beliefs for *Ind-NT* and group prior beliefs for *Group-NT*. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

4.3 Transfers

Figure 2 presents the average transfers made by individuals and groups in the trust game. We find no significant difference in the average transfers made by individuals and groups in Group A (*Ind*: 27.97 vs. *Group*: 30.00, $p = 1.00$), and in Group B (18.67 vs. 16.73, $p = 0.40$). These results are also reflected in the regression analysis reported in Table 6. Following communication, Group A members do not send a significantly different amount ($p = 0.75$, column 1), as compared to individuals representing their groups. For Group B, groups appear to send less than individuals, however this is also not significant ($p = 0.60$, column 2).

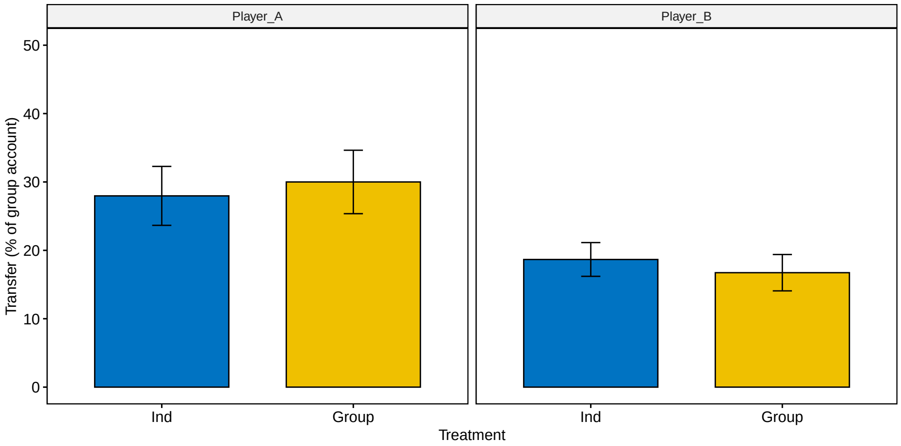
Consistent with findings from previous work, Table 6 shows a significantly positive relationship between prior beliefs and transfers made by both individuals ($p < 0.01$, column

Table 5: Belief updating by individuals and groups

	Ind	*Group*
	(1)	(2)
δ^{prior}	0.29**	0.31**
	(0.12)	(0.14)
β^{high}	−0.17	−0.21
	(0.22)	(0.21)
β^{low}	0.87*	0.78***
	(0.46)	(0.27)
δ^{prior} x Transfer	0.25	−0.08
	(0.20)	(0.16)
β^{high} x Transfer	0.38	−0.11
	(0.34)	(0.29)
β^{low} x Transfer	−0.44	0.09
	(0.52)	(0.30)
$H_0 : \delta^{prior} = 1$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{high} = 1$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p = 0.77$	$p = 0.41$
$H_0 : \beta^{high} = \beta^{low}$	$p = 0.07$	$p = 0.02$
R^2	0.30	0.41
Adj. R^2	0.30	0.41
Num. obs.	540	483

Notes: Ordinary least squares regression with standard errors clustered at the individual level (in *Ind*) and at the group level (in *Group*) in parentheses. The dependent variable is the belief about the proportion of groups that send a quarter or more of their account. We use individual prior beliefs for individuals and group prior beliefs for groups. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Figure 2: Transfers by individuals and groups



Note: Error bars represent standard errors.

Table 6: Transfers by individuals and groups (as a % of group account)

	Group A (1)	Group B (2)	Pooled (3)
Group	5.38	-1.82	1.32
	(6.04)	(3.50)	(3.61)
Prior belief	0.37^{***}	0.16^{**}	0.26^{***}
	(0.12)	(0.07)	(0.07)
Constant	13.57^{**}	12.93^{***}	13.69^{***}
	(6.37)	(3.49)	(3.72)
R^2	0.14	0.09	0.10
Adj. R^2	0.11	0.06	0.08
Num. obs.	60	60	120

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the transfer made (as a % of the group account). The prior belief is the individual prior for *Ind* and the group prior in *Group*. In *Group*, we exclude 4 Group A's and 6 Group B's that failed to reach an agreement on the group prior. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

1) and groups ($p < 0.01$, columm 2). This pattern holds when we pool data from both the *Ind* and *Group* treatments ($p < 0.01$, column 3), indicating that transfers are increasing in participants' beliefs about the likelihood of being matched with a group that transferred a relatively greater amount. In Appendix G, we show that individuals and groups with relatively lower prior beliefs transfer less than those with higher priors.

Result 4: We do not find a significant difference between group transfers and the transfers made by individuals representing their groups. Transfers are, however, increasing in both individual and group prior beliefs.

4.4 Text analysis of communication data

To further explore the role of communication in group decisions (Cooper and Kagel, 2005), we perform text analyses on Group B chat data.[27] Our goal is to better understand the motivations and reasons used to justify belief and transfer decisions. We organize the most frequently-used relevant keywords into four main topics.[28] The topics Low and High contain keywords which specify magnitude or relative magnitude related to the signal, belief or transfer amount, i.e., "less", "low", "small", "greater", "high" and "large". The topic Fairness

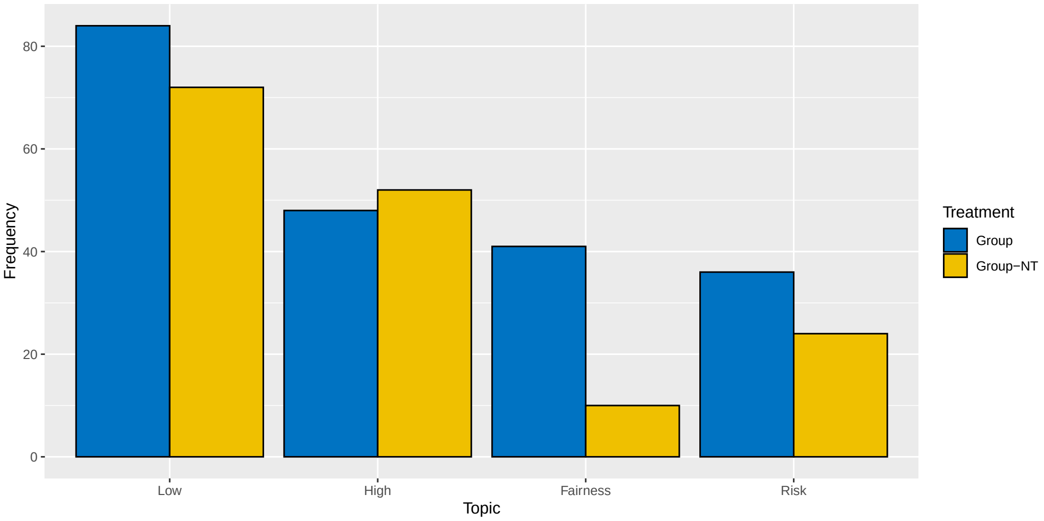
[27] We focus on Group B's given that our main analysis is on the effect of communication on belief updating.

[28] We excluded stop words, punctuation, symbols and "stemmed" words which share the same root, e.g., "low" and "lower" would be grouped together. Non-relevant keywords include words that are not related to opinions about beliefs or transfers, e.g., "think", "okay", "yeah", etc.

offers insight into motivations related to fairness and selfishness, including the words “nice”, “fair”, “generous”, “greed”, “selfish” and “profit”. One example of such a message is: “I doubt people will be so generous”. The topic Risk also includes keywords that explain participants’ decisions and highlights the risky nature of the transfer decision, including “trust”, “lose”, “safe”, “risk” and “faith”. For example, one message containing the keyword “risk” is: “I feel there will still be some groups that just wanna [sic] take the 100 and not risk anything”.

Figure 3 depicts the distribution of topics across the Transfer and No-Transfer treatments. Groups in the Transfer treatment are more likely to use words related to Fairness and Risk, compared to those in the No-Transfer treatment. According to a χ^2 test, the distribution of topics is significantly different across treatments ($p < 0.01$). Qualitatively when we examine messages related to Risk, third parties are more likely to highlight the low-risk nature of the decision, e.g., “I think should be quite high because it’s low risk”, while second parties are more likely to characterise the decision as being higher risk, e.g., “the more they send to us they more they will lose”.

Figure 3: Distribution of topics in the Transfer and No-Transfer treatments

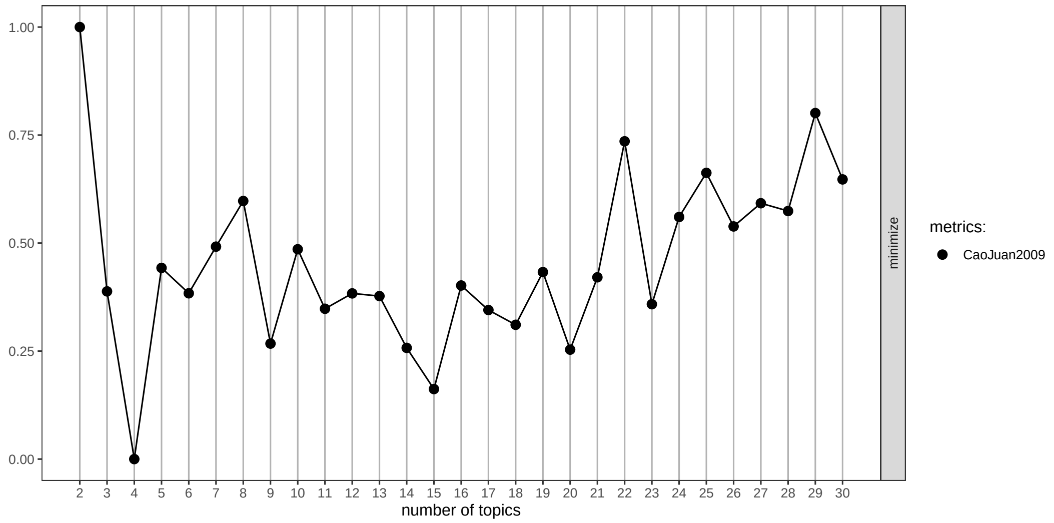


We find similar results when we use a topic modeling approach to uncover underlying topics in communication. Following Hanaki and Ozkes (2023) and Andres et al. (2022), we use Latent Dirichlet Allocation (LDA), an unsupervised machine learning algorithm which finds natural groups of text data, see Blei (2012) for an overview of the procedure. One major advantage of this approach is that it does not rely on subjective assessments by researchers in predefining categories.[29] We first organize the chat data into “documents”, one for each

[29] See Brandts et al. (2019) and Hanaki and Ozkes (2023) for discussions of different methods of classifying

line of chat messages sent by Group B members.[30] The text corpus is thus made up of 3950 documents. To minimize noise in the LDA model, we remove stopwords, correct for spelling mistakes, coerce all words to lower case, remove numbers and punctuation, and allow each token (i.e., keywords or key phrases) to be between 1-2 words long.[31] We use the method by Cao et al. (2009) which adaptively selects the optimal number of topics based on topic density. Figure 4 presents the Cao et al. (2009) metric which is minimized for the optimal LDA model. Based on this, we estimate a LDA model with $K = 15$ topics which minimizes the Cao et al. (2009) metric for a sufficient number of topics. Therefore, both the number of topics and the grouping of topics are selected by the algorithm based on patterns within the chat data.

Figure 4: Selecting the optimal number of topics



Figures 5-8 presents the top ten tokens for Topics 6, 7, 10 and 12 and the probability of observing each token in the chat data (see Appendix H for all 15 topics). A common feature of these four topics is that they contain keywords related to risk in the top ten list, i.e., “trust”, “safe”, “lose”, “faith” and “risk”. Examples of such messages include: “group A loses least if they put 1%”, “0 is the only safe line”, and “yeah that seems fair, i [sic] guess especially in groups where they may have been more likely to be risk averse?”. The results of the LDA analysis therefore support the findings of the text analysis that risk preferences play a key role in group communication and groups’ decisions.

The machine learning approach also offers support for the role of social preferences in

free-form communication.

[30] See Appendix H for the LDA model for Group A chat data.

[31] Examples of stopwords include: “the”, “a” and “are”. Examples of 2-word tokens include: “send back”, “group transfer” and “fairly low”.

Figure 5: Topic 6

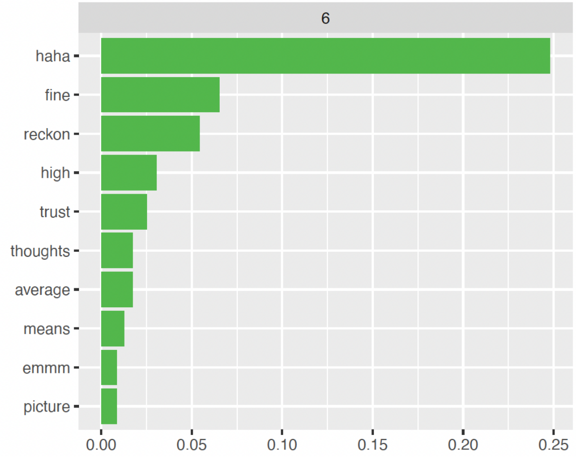


Figure 6: Topic 7

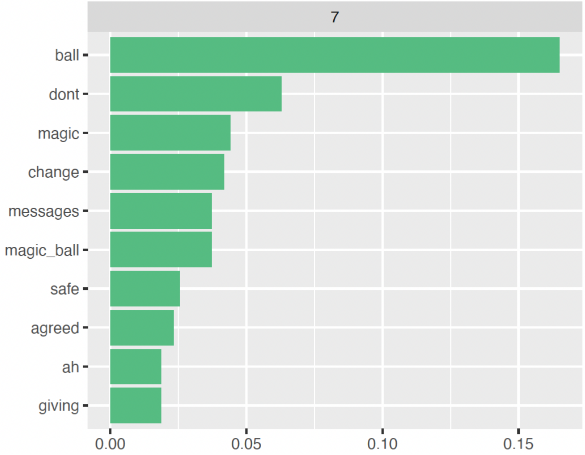


Figure 7: Topic 10

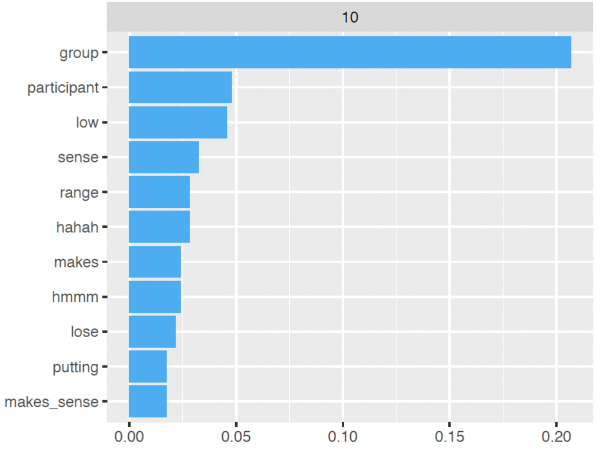
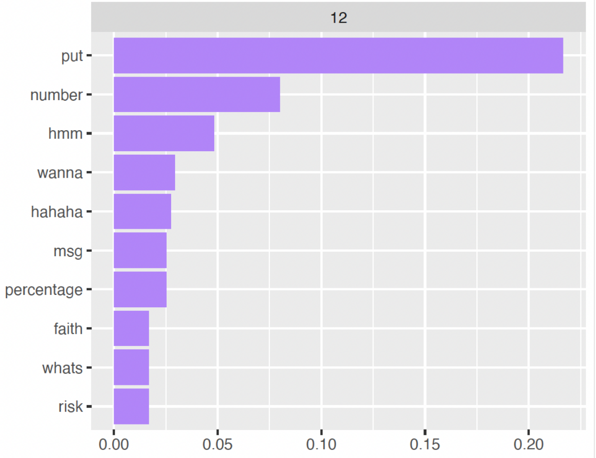


Figure 8: Topic 12



players' decisions, consistent with the existing literature on the trust game. Topics 3 and 11 (Figures 9 and 10) features keywords related to fairness in the top ten, i.e., "greedy" and "generous", for example: "let's be greedy", "I'm feeling generous", and "I doubt people will be so generous".

4.5 Risk preferences

Based on the text analysis of communication, risk appears to play a major role in group beliefs and choices in the trust game. We subsequently explore the role of risk attitudes in transfer decisions as well as in prior and posterior beliefs.[32]

[32] We show in Appendix I that individual risk preferences do not differ across the Transfer and No-Transfer treatments, nor between *Ind* and *Group*.

Figure 9: Topic 3

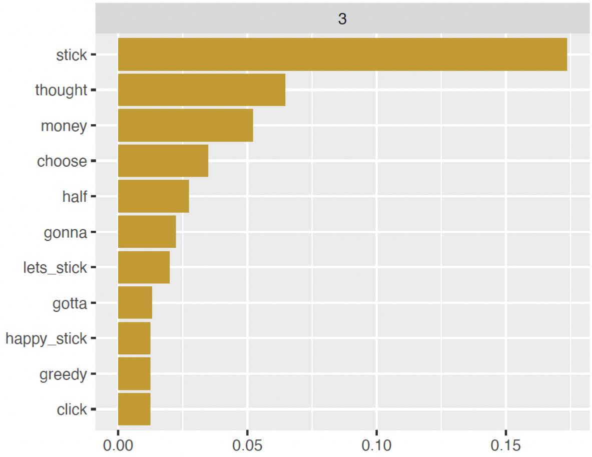
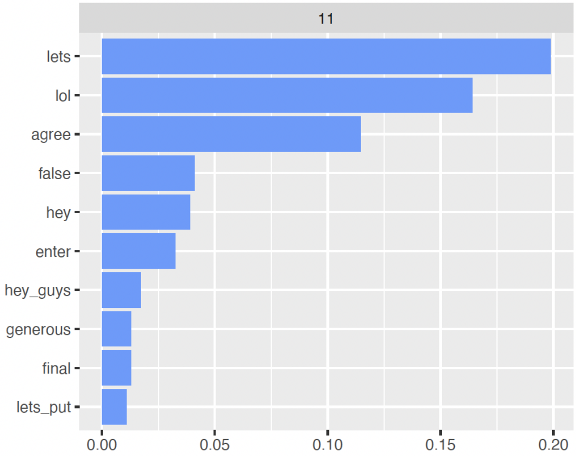


Figure 10: Topic 11



4.5.1 Risk preferences and transfers

We use the Eckel and Grossman (2002) measure to classify individuals and groups as relatively more risk-averse, or relatively more risk-tolerant. We classify individuals as being more risk-averse if they chose one of the less risky lotteries (lotteries 1, 2, or 3), while those who chose one of the more risky lotteries (lotteries 4, 5, or 6) are classified as more risk-tolerant.[33] In *Group*, we use the number of risk-averse group members as a measure of the group's risk profile.

Table 7 presents the relationship between risk attitudes and transfers made in the Transfer treatment. In *Ind*, we examine the relationship between transfer choices and individual risk preferences. We do not observe a significant relationship for either Group A ($p = 0.96$, column 1), Group B ($p = 0.86$, column 2) or when we pool the two groups ($p = 0.95$, column 3). We find some evidence that Group A's transfers are negatively correlated with the number of risk-averse members in the group ($p = 0.07$, column 4). The direction of the effect is also negative for Group B's ($p = 0.78$, column 5) and when we pool data for both groups ($p = 0.28$, column 6), though not significantly so.

4.5.2 Risk preferences and prior beliefs

We next examine whether a relationship exists between risk preferences and prior beliefs for individuals and groups. Table 8 shows that the more risk-tolerant an individual is, the greater their individual prior belief and this is significant for Group B ($p = 0.04$, column 2) and the pooled data ($p = 0.03$, column 3), though not for Group A ($p = 0.19$, column 1).

[33] This classification is intended to approximately divide participants into two groups – one that is relatively more risk-averse and one that is relatively more risk-tolerant – rather than as an label of absolute risk-aversion.

Table 7: Group and individual transfers by risk attitudes

	Ind			*Group*		
	Group A	Group B	Pooled	Group A	Group B	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Ind Risk	0.12	0.37	0.10			
	(2.54)	(2.04)	(1.70)			
# Risk-averse				-11.47^{*}	-1.01	-4.04
				(6.12)	(3.55)	(3.67)
Constant	27.56^{***}	17.36^{**}	22.97^{***}	49.88^{***}	18.18^{***}	29.77^{***}
	(9.83)	(7.70)	(6.52)	(11.50)	(5.76)	(6.44)
R^2	0.00	0.00	0.00	0.11	0.00	0.02
Adj. R^2	-0.04	-0.03	-0.02	0.08	-0.03	0.00
Num. obs.	30	30	60	30	30	60

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the transfer made (as a % of the group account) by the individual for columns 1-3 and by the group for columns 4-6. # Risk-averse takes a value between 0 and 3 for the number of group members who are classified as more risk-averse. Ind Risk denotes the lottery choice between 1 and 6. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

As the number of risk-averse group members increases, the group prior tends to decrease for Group A ($p = 0.03$, column 4), Group B ($p = 0.08$, column 5) and when we pool the two types of groups ($p < 0.01$, column 6). This is largely consistent with the findings for individual priors.

4.5.3 Belief updating by group risk profile

Given the correlation between risk preferences and prior beliefs, we investigate whether risk also plays a role in belief updating.[34] Columns 1 and 2 of Table 9 compare belief updating across *Group* and *Group-NT* based on the group risk profile. We classify groups as "risk-averse" if a majority (two or more members) are classified as more risk-averse, and "risk-tolerant" if two or more members are more risk-tolerant. In the No-Transfer treatment, risk-averse groups place more weight on prior beliefs than risk-tolerant groups ($p = 0.05$, column 2), but place similar weights on high signals ($p = 0.42$) and low signals ($p = 0.27$). In the Transfer treatment, however, the two group types do not differ in the weight on prior beliefs ($p = 0.24$, column 1), but risk-averse groups place substantially less weight on high signals ($p < 0.01$) and more weight on low signals ($p = 0.02$), which is indicative of a self-serving bias in updating.

[34] Appendix J examines belief updating based on whether groups transferred less or more than the median transfer, as a proxy for social preferences.

Table 8: Group and individual priors by risk attitudes

	Individual prior			Group prior		
	Group A	Group B	Pooled	Group A	Group B	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Ind Risk	1.38	2.99^{**}	1.86^{**}			
	(1.06)	(1.48)	(0.86)			
# Risk-averse				-9.17^{**}	-10.57^{*}	-9.55^{***}
				(4.19)	(5.88)	(3.35)
Constant	35.10^{***}	27.67^{***}	32.88^{***}	48.27^{***}	49.35^{***}	48.56^{***}
	(4.03)	(5.83)	(3.31)	(7.58)	(9.55)	(5.86)
R^2	0.01	0.03	0.01	0.08	0.10	0.08
Adj. R^2	0.00	0.02	0.01	0.06	0.07	0.07
Num. obs.	240	120	360	60	30	90

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the individual prior for columns 1-3 and the group prior for columns 4-6. # Risk-averse takes a value between 0 and 3 for the number of group members who are classified as more risk-averse. Ind Risk denotes the lottery choice between 1 and 6. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

We then compare, for each group risk profile, whether updating differs across the Transfer and No-Transfer treatments, which would indicate a self-serving bias. Groups that are more risk-averse in the Transfer treatment place less weight on prior beliefs ($p < 0.01$, column 3), less weight on high signals ($p < 0.01$) and more weight on low signals ($p = 0.04$), as compared to third party groups in the No-Transfer treatment. Risk-averse groups are, therefore, more likely to be motivated to believe the other group did not transfer much. However, we do not observe this in more risk-tolerant groups, as these groups place similar weights on prior beliefs ($p = 0.45$, column 4), high signals ($p = 0.98$) and low signals ($p = 0.52$) in the Transfer and No-Transfer treatments.

Thus, while we find no evidence of a self-serving bias for groups overall, our findings suggest that Result 3 is masking some important heterogeneity in belief updating, based on the risk appetite of groups. We find that risk attitudes matter for how groups process new information, with greater risk aversion associated with more biased belief updating.

4.5.4 Belief updating by individual risk attitudes

We next investigate whether risk attitudes also affect belief updating at the individual level. While there is some evidence that more risk-averse individuals place less weight on high signals ($p = 0.07$, column 1, Table 10) than their more risk-tolerant counterparts in *Ind*, this is also the case in *Ind-NT* ($p = 0.08$, column 2).

Table 9: Belief updating in *Group* by group risk profile

	Group	*Group-NT*	Risk-averse	Risk-tolerant
	(1)	(2)	(3)	(4)
δ^{prior}	0.24^{***}	0.11	0.42^{***}	0.12
	(0.04)	(0.14)	(0.07)	(0.16)
β^{high}	-0.06	-0.07	-0.26^{*}	-0.05
	(0.14)	(0.19)	(0.14)	(0.21)
β^{low}	0.68^{***}	0.44	0.79^{***}	0.43
	(0.22)	(0.27)	(0.16)	(0.30)
δ^{prior} x Risk-averse	-0.08	0.31^{*}		
	(0.07)	(0.16)		
β^{high} x Risk-averse	-0.78^{***}	-0.19		
	(0.21)	(0.23)		
β^{low} x Risk-averse	0.63^{**}	0.35		
	(0.27)	(0.32)		
δ^{prior} x Transfer			-0.29^{***}	0.13
			(0.09)	(0.17)
β^{high} x Transfer			-0.55^{***}	-0.01
			(0.20)	(0.25)
β^{low} x Transfer			0.45^{**}	0.24
			(0.22)	(0.38)
$H_0 : \delta^{prior} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{high} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p = 0.14$	$p = 0.04$	$p = 0.18$	$p = 0.06$
$H_0 : \beta^{high} = \beta^{low}$	$p < 0.01$	$p = 0.12$	$p < 0.01$	$p = 0.17$
R^2	0.53	0.39	0.65	0.14
Adj. R^2	0.51	0.37	0.64	0.12
Num. obs.	249	234	261	222

Notes: Ordinary least squares regression with standard errors clustered at the group level in parentheses. The dependent variable is the group belief about the proportion of groups that send a quarter or more of their account. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

However, in contrast to groups, we do not find a significant difference between the Transfer and No-Transfer treatments among risk-averse individuals for either high signals ($p = 0.75$, column 3) or low signals ($p = 0.41$). Our results are similar for risk-tolerant individuals for high signals ($p = 0.90$, column 4) and low signals ($p = 0.31$). Individuals do not appear to update their beliefs in a self-serving way, regardless of their risk appetite.

Taken together, we find evidence of a self-serving bias in belief updating by groups that are more risk-averse. However, we do not observe more biased updating by risk-averse individuals. Our results suggest that risk aversion and decision making in groups on their own are not sufficient to generate motivated biases in updating, however, the *interaction*

Table 10: Belief updating in *Ind* by individual risk attitudes

	Ind	*Ind-NT*	Risk-averse	Risk-tolerant
	(1)	(2)	(3)	(4)
δ^{prior}	0.67***	0.41***	0.37**	0.26
	(0.15)	(0.15)	(0.17)	(0.31)
β^{high}	0.52	−0.00	−0.10	−0.33
	(0.37)	(0.28)	(0.30)	(0.33)
β^{low}	0.33	0.87	0.77*	0.74*
	(0.45)	(0.53)	(0.40)	(0.43)
δ^{prior} x Risk-averse	−0.29	−0.50*		
	(0.22)	(0.27)		
β^{high} x Risk-averse	−0.74*	−0.67*		
	(0.43)	(0.38)		
β^{low} x Risk-averse	0.37	−0.25		
	(0.54)	(0.69)		
δ^{prior} x Transfer			−0.10	−0.09
			(0.23)	(0.31)
β^{high} x Transfer			−0.14	−0.09
			(0.50)	(0.40)
β^{low} x Transfer			0.27	0.06
			(0.55)	(0.45)
$H_0 : \delta^{prior} = 1$	$p = 0.03$	$p < 0.01$	$p < 0.01$	$p = 0.02$
$H_0 : \beta^{high} = 1$	$p = 0.19$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p = 0.14$	$p = 0.80$	$p = 0.56$	$p = 0.55$
$H_0 : \beta^{high} = \beta^{low}$	$p = 0.79$	$p = 0.20$	$p = 0.15$	$p = 0.12$
R^2	0.50	0.24	0.56	0.24
Adj. R^2	0.49	0.22	0.55	0.22
Num. obs.	261	270	234	249

Notes: Ordinary least squares regression with standard errors clustered at the individual level in parentheses. The dependent variable is the individual belief about the proportion of groups that send a quarter or more of their account. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

between the two does lead to more biased beliefs. While we do not find a tendency toward motivated reasoning for all groups, the risk profile of a group is important in determining how new information is incorporated into beliefs.

5 Conclusion

In this paper, we examine the role of communication in group belief updating in the presence of a self-interested motive to bias beliefs. We directly elicit the beliefs of groups (that are required to reach a consensus) and compare these to the beliefs of individual group members, thus isolating within-group communication in the belief formation process. We experimentally manipulate the presence of a self-serving motive and compare the beliefs of participants who are motivated to minimize the cognitive dissonance between maximizing payoffs and believing their actions to be fair, and disinterested third parties, who do not have a vested interest in the outcome. We conjecture that groups are more likely to hold biased beliefs because the psychological cost of distorting beliefs is lower for groups following communication.

Our main findings are as follows. Overall, groups do not differ significantly from individuals in transfer decisions or prior beliefs. Groups exhibit asymmetric updating (and are more responsive to low signals than high signals) but do not appear to be more biased than individuals in belief updating. However, we do observe a self-serving bias by groups that are composed of more risk-averse members, while this is not present in groups with more risk-tolerant members. Crucially, individuals who are more risk-averse do not exhibit such biases. Taken together, our results highlight the role of the group decision-making environment and group composition in intensifying individual risk preferences, with consequences for how new information is incorporated into group beliefs.

Our research has several policy implications. First, we find no evidence to suggest that communication and the group decision-making environment *per se* contribute to more biased information processing. While many major decisions are already made by committees, our results do not suggest that group decision making is necessarily worse than individual decisions. However, a second important implication is the potential for the group environment to exaggerate individual members' characteristics. This is consistent with conclusions by Kerr et al. (1996) that groups can attenuate, amplify or reproduce individual biases and that the effect of group discussion depends on the decision making context. We contribute to the literature examining whether and how group decisions differ from individual decisions by offering direct evidence on belief updating following within-group communication. Our findings highlight the need for a greater diversity of perspectives within organizations in order to minimize the potential for biased decision making. Finally, we contribute to a small but growing literature that do not find evidence of excuse-driven selfishness (e.g., Van der Weele et al., 2014; Bartling and Özdemir, 2023; Valero, 2021) and that more work is needed to better understand the conditions that facilitate motivated reasoning. Drobner (2022) offers evidence that motivated reasoning is less likely when subjects expect to receive feedback straightaway. Both Ging-Jehli et al. (2020) and Gangadharan et al. (2022) elicit beliefs from neutral observers and find no evidence that beliefs are distorted to excuse selfish decisions. While there is growing interest in documenting instances of motivated reasoning, more work is needed to better understand *when* beliefs are more likely to be distorted. Our results suggest that both communication and the composition of a group matter in providing a fertile environment for biased beliefs.

An interesting avenue for future work is to examine whether communication in groups can amplify other individual preferences, attitudes or biases that could also lead to biased beliefs. Where the composition of groups cannot be altered or is difficult to change, another important area for future research is to investigate how biased belief updating in groups could be mitigated through other means.

References

Akerlof, G. A. and Dickens, W. T. (1982). The economic consequences of cognitive dissonance. *The American Economic Review*, 72(3):307–319.

Ambrus, A., Greiner, B., and Pathak, P. A. (2015). How individual preferences are aggregated in groups: An experimental study. *Journal of Public Economics*, 129:1–13.

Andres, M., Bruttel, L. V., and Friedrichsen, J. (2022). How communication makes the difference between a cartel and tacit collusion.

Ashraf, N., Bohnet, I., and Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics*, 9(3):193–208.

Babcock, L., Loewenstein, G., Issacharoff, S., and Camerer, C. (1995). Biased judgments of fairness in bargaining. *The American Economic Review*, 85(5):1337–1343.

Bainbridge, S. M. (2002). Why a board-group decisionmaking in corporate governance. *Vand. L. Rev.*, 55:1.

Bartling, B. and Özdemir, Y. (2023). The limits to moral erosion in markets: Social norms and the replacement excuse. *Games and Economic Behavior*, 138:143–160.

Behnk, S., Hao, L., and Reuben, E. (2022). Shifting normative beliefs: On why groups behave more antisocially than individuals. *European Economic Review*, 145:104116.

Bénabou, R. (2013). Groupthink: Collective delusions in organizations and markets. *Review of Economic Studies*, 80(2):429–462.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, 2:69–186.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms.* Cambridge University Press.

Bicchieri, C., Dimant, E., and Sonderegger, S. (2023). It's not a lie if you believe the norm does not apply: Conditional norm-following and belief distortion. *Games and Economic Behavior*, 138:321–354.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bohnet, I. and Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior and Organization*, 55(4):467–484.

Bornstein, G., Kugler, T., and Ziegelmeyer, A. (2004). Individual and group decisions in the centipede game: Are groups more "rational" players? *Journal of Experimental Social Psychology*, 40(5):599–605.

Bornstein, G. and Yaniv, I. (1998). Individual and group behavior in the ultimatum game: are groups more "rational" players? *Experimental Economics*, 1(1):101–108.

Brandts, J., Cooper, D. J., and Rott, C. (2019). 21. communication in laboratory experiments. *Handbook of research methods and applications in experimental economics*, 401.

Campbell, D. T. (1967). Stereotypes and the perception of group differences. *American psychologist*, 22(10):817.

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7-9):1775–1781.

Cason, T. N. and Mui, V.-L. (1997). A laboratory study of group polarisation in the team dictator game. *The Economic Journal*, 107(444):1465–1483.

Castagnetti, A. and Schmacker, R. (2022). Protecting the ego: Motivated information selection and updating. *European Economic Review*, 142:104007.

Charness, G., Gneezy, U., and Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior and Organization*, 189:234–256.

Charness, G., Karni, E., and Levin, D. (2007). Individual and group decision making under risk: An experimental study of bayesian updating and violations of first-order stochastic dominance. *Journal of Risk and Uncertainty*, 35(2):129–148.

Charness, G. and Sutter, M. (2012). Groups make better self-interested decisions. *Journal of Economic Perspectives*, 26(3):157–76.

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

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–457.

Chew, S. H., Huang, W., and Zhao, X. (2020). Motivated false memory. *Journal of Political Economy*, 128(10):3913–3939.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6):1015.

Cooper, D. J. and Kagel, J. H. (2005). Are two heads better than one? Team versus individual play in signaling games. *American Economic Review*, 95(3):477–509.

Costa-Gomes, M. A., Huck, S., and Weizsäcker, G. (2014). Beliefs and actions in the trust game: Creating instrumental variables to estimate the causal effect. *Games and Economic Behavior*, 88:298–309.

Coutts, A. (2019). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics*, 22(2):369–395.

Cox, C. A. and Stoddard, B. (2018). Strategic thinking in public goods games with teams. *Journal of Public Economics*, 161:31–43.

Cox, J. C. (2002). Trust, reciprocity, and other-regarding preferences: Groups vs. individuals and males vs. females. In *Experimental business research*, pages 331–350. Springer.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–74.

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015). Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism. *American Economic Review*, 105(11):3416–42.

Drobner, C. (2022). Motivated beliefs and anticipation of uncertainty resolution. *American Economic Review: Insights*, 4(1):89–105.

Eckel, C. C. and Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, 23(4):281–295.

Eckel, C. C. and Grossman, P. J. (2005). Managing diversity by creating team identity. *Journal of Economic Behavior and Organization*, 58(3):371–392.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–138.

Erkal, N., Gangadharan, L., and Koh, B. H. (2021). By chance or by choice? Biased attribution of others' outcomes when social preferences matter. *Experimental Economics*, pages 1–31.

Fahr, R. and Irlenbusch, B. (2011). Who follows the crowd—groups or individuals? *Journal of Economic Behavior and Organization*, 80(1):200–209.

Fehr, E. (2009). On the economics and biology of trust. *Journal of the European Economic Association*, 7(2-3):235–266.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Gangadharan, L., Grossman, P. J., Xue, N., et al. (2022). Belief elicitation under competing motivations: Does it matter how you ask? *Monash University Department of Economics Working Paper*.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Ging-Jehli, N. R., Schneider, F. H., and Weber, R. A. (2020). On self-serving strategic beliefs. *Games and Economic Behavior*, 122:341–353.

Gino, F., Norton, M. I., and Weber, R. A. (2016). Motivated bayesians: Feeling moral while acting egoistically. *Journal of Economic Perspectives*, 30(3):189–212.

Glassop, L. I. (2002). The organizational benefits of teams. *Human Relations*, 55(2):225–249.

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

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics*, 95(3):537–557.

Guerra, G. and Zizzo, D. J. (2004). Trust responsiveness and beliefs. *Journal of Economic Behavior and Organization*, 55(1):25–30.

Hanaki, N. and Ozkes, A. I. (2023). Strategic environment effect and communication. *Experimental Economics*, 26(3):588–621.

Houser, D., Schunk, D., and Winter, J. (2010). Distinguishing trust from risk: An anatomy of the investment game. *Journal of Economic Behavior and Organization*, 74(1-2):72–81.

Insko, C. A. and Schopler, J. (1987). Categorization, competition, and collectivity.

Janis, I. L. (1972). *Victims of groupthink*. Boston: Houghton-Mifflin.

Janis, I. L. (1983). *Groupthink*. Boston: Houghton-Mifflin.

Karlan, D. S. (2005). Using experimental economics to measure social capital and predict financial decisions. *American Economic Review*, 95(5):1688–1699.

Kerr, N. L., MacCoun, R. J., and Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological review*, 103(4):687.

Kocher, M. G., Praxmarer, M., and Sutter, M. (2020). Team decision-making. *Handbook of Labor, Human Resources and Population Economics*, pages 1–25.

Kocher, M. G., Schudy, S., and Spantig, L. (2018). I lie? we lie! why? experimental evidence on a dishonesty shift in groups. *Management Science*, 64(9):3995–4008.

Kocher, M. G. and Sutter, M. (2005). The decision maker matters: Individual versus group behaviour in experimental beauty-contest games. *The Economic Journal*, 115(500):200–223.

Kogan, S., Schneider, F., and Weber, R. (2021). Self-serving biases in beliefs about collective outcomes.

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review*, 90(4):1072–1091.

Kugler, T., Bornstein, G., Kocher, M. G., and Sutter, M. (2007). Trust between individuals and groups: Groups are less trusting than individuals but just as trustworthy. *Journal of Economic Psychology*, 28(6):646–657.

Kugler, T., Kausel, E. E., and Kocher, M. G. (2012). Are groups more rational than individuals? A review of interactive decision making in groups. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(4):471–482.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3):480.

Luhan, W. J., Kocher, M. G., and Sutter, M. (2009). Group polarization in the team dictator game reconsidered. *Experimental Economics*, 12(1):26–41.

Maciejovsky, B., Sutter, M., Budescu, D. V., and Bernau, P. (2013). Teams make you smarter: How exposure to teams improves individual decisions in probability and reasoning tasks. *Management Science*, 59(6):1255–1270.

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, 69(3):220.

Mengel, F. (2021). Gender bias in opinion aggregation. *International Economic Review*, 62(3):1055–1080.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2022). Managing self-confidence. *Management Science*.

Naef, M., Fehr, E., Fischbacher, U., Schupp, J., and Wagner, G. G. (2008). Decomposing trust: Explaining national differences. *International Journal of Psychology*, 43(3-4):577–577.

Oprea, R. and Yuksel, S. (2020). Social exchange of motivated beliefs. *Journal of the European Economic Association*.

Penczynski, S. P. (2019). Using machine learning for communication classification. *Experimental Economics*, 22(4):1002–1029.

Rabin, M. (1994). Cognitive dissonance and social change. *Journal of Economic Behavior and Organization*, 23(2):177–194.

Schechter, L. (2007). Traditional trust measurement and the risk confound: An experiment in rural Paraguay. *Journal of Economic Behavior and Organization*, 62(2):272–292.

Schlag, K. and Tremewan, J. (2021). Simple belief elicitation: An experimental evaluation. *Journal of Risk and Uncertainty*, 62:137–155.

Schopler, J. and Insko, C. A. (1992). The discontinuity effect in interpersonal and intergroup relations: Generality and mediation. *European review of social psychology*, 3(1):121–151.

Song, F. (2008). Trust and reciprocity behavior and behavioral forecasts: Individuals versus group-representatives. *Games and Economic Behavior*, 62(2):675–696.

Sutter, M., Czermak, S., and Feri, F. (2013). Strategic sophistication of individuals and teams. experimental evidence. *European Economic Review*, 64:395–410.

Tajfel, H. and Turner, J. C. (1979). An integrative theory of intergroup conflict. In Austin, W. G. and Worchel, S., editors, *The Social Psychology of Intergroup Relations*, pages 33–47. Monterey, CA:Brooks/Cole.

Valero, V. (2021). Redistribution and beliefs about the source of income inequality. *Experimental Economics*, pages 1–26.

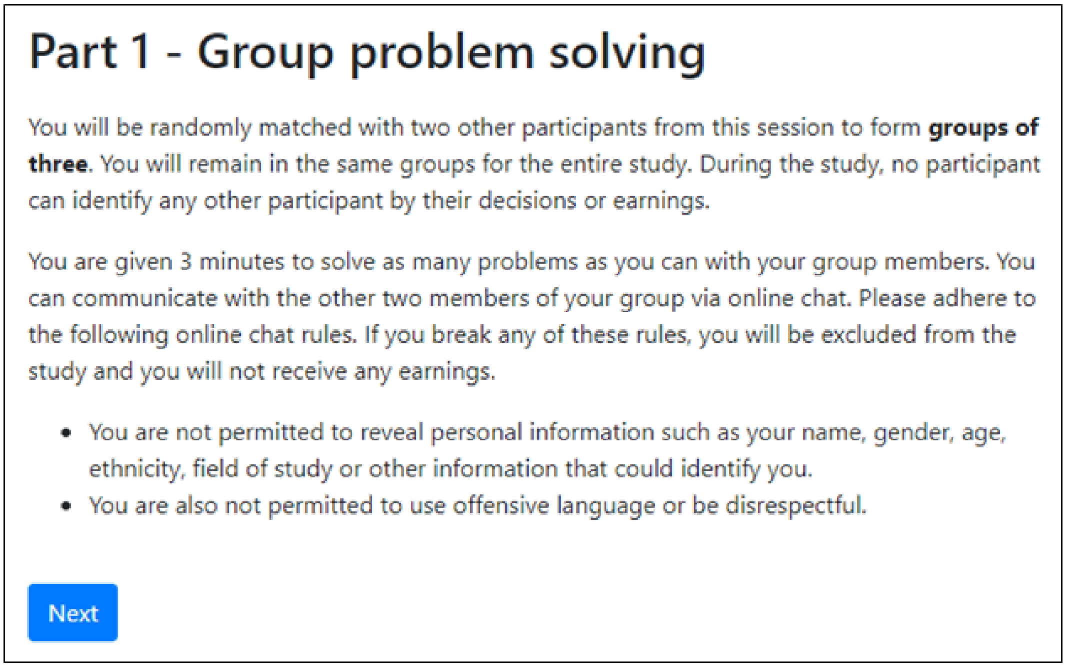
Van der Weele, J. J., Kulisa, J., Kosfeld, M., and Friebel, G. (2014). Resisting moral wiggle room: how robust is reciprocal behavior? *American Economic Journal: Microeconomics*, 6(3):256–64.

Winquist, J. R. and Larson, J. R. J. (1998). Information pooling: When it impacts group decision making. *Journal of personality and social psychology*, 74(2):371.

Zimmermann, F. (2020). The dynamics of motivated beliefs. *American Economic Review*, 110(2):337–61.

A Instructions

Figure A.1: Problem solving task instructions



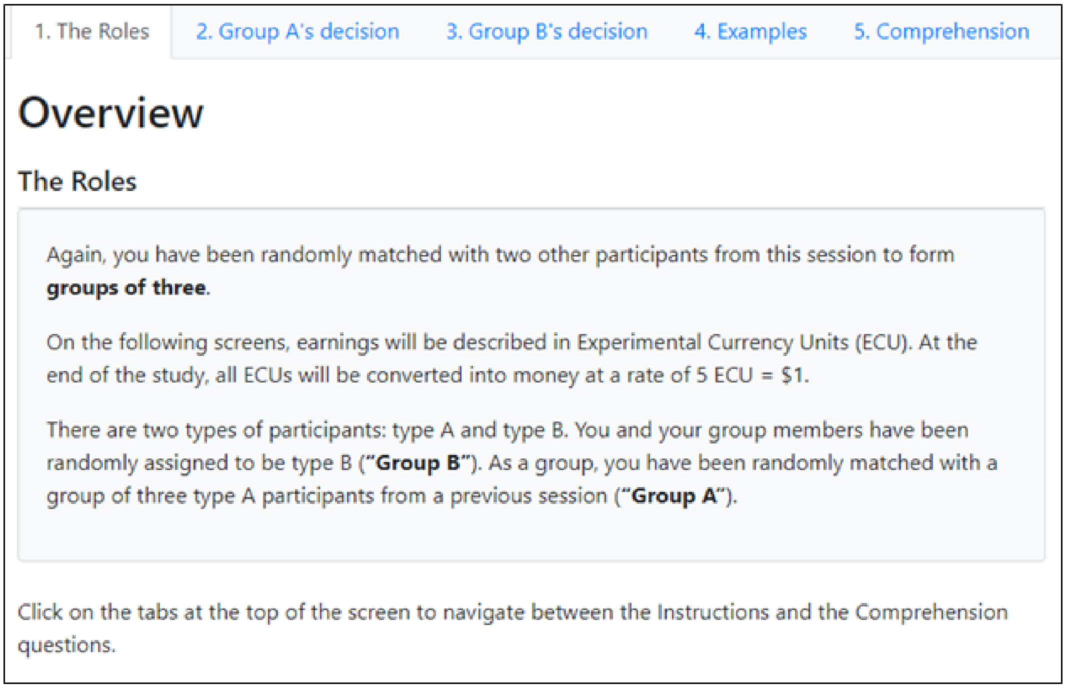
Part 1 - Group problem solving

You will be randomly matched with two other participants from this session to form **groups of three**. You will remain in the same groups for the entire study. During the study, no participant can identify any other participant by their decisions or earnings.

You are given 3 minutes to solve as many problems as you can with your group members. You can communicate with the other two members of your group via online chat. Please adhere to the following online chat rules. If you break any of these rules, you will be excluded from the study and you will not receive any earnings.

- You are not permitted to reveal personal information such as your name, gender, age, ethnicity, field of study or other information that could identify you.
- You are also not permitted to use offensive language or be disrespectful.

Figure A.2: Overview of instructions (Group B)



Overview

The Roles

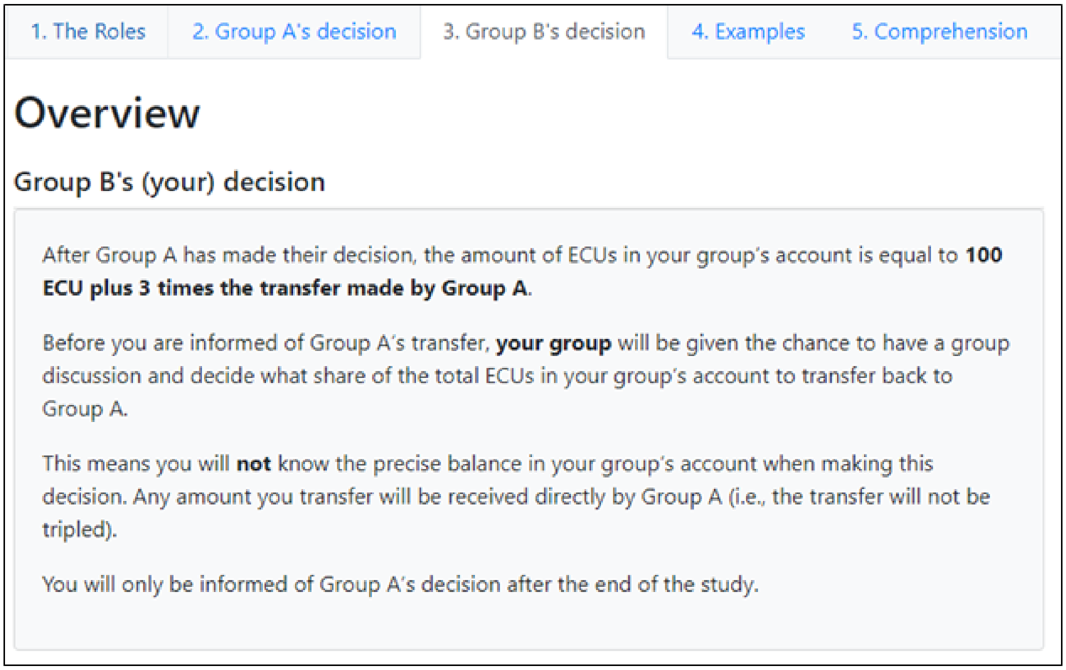
Again, you have been randomly matched with two other participants from this session to form **groups of three**.

On the following screens, earnings will be described in Experimental Currency Units (ECU). At the end of the study, all ECUs will be converted into money at a rate of 5 ECU = $1.

There are two types of participants: type A and type B. You and your group members have been randomly assigned to be type B (**"Group B"**). As a group, you have been randomly matched with a group of three type A participants from a previous session (**"Group A"**).

Click on the tabs at the top of the screen to navigate between the Instructions and the Comprehension questions.

Figure A.3: Overview of instructions (Group B)



Overview

Group B's (your) decision

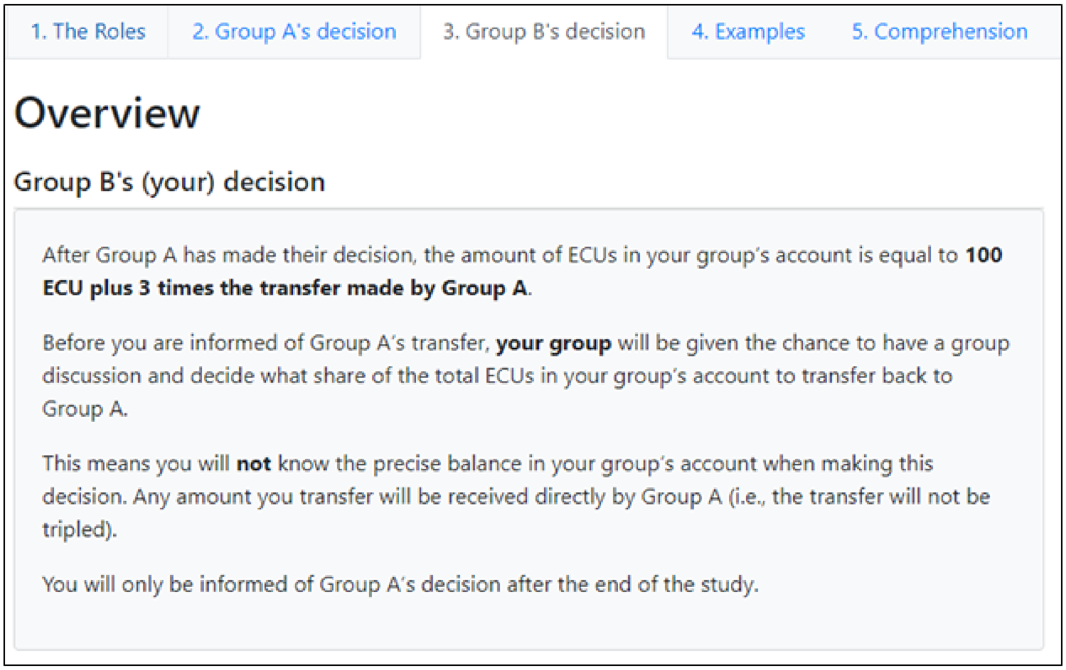
After Group A has made their decision, the amount of ECUs in your group's account is equal to **100 ECU plus 3 times the transfer made by Group A**.

Before you are informed of Group A's transfer, **your group** will be given the chance to have a group discussion and decide what share of the total ECUs in your group's account to transfer back to Group A.

This means you will **not** know the precise balance in your group's account when making this decision. Any amount you transfer will be received directly by Group A (i.e., the transfer will not be tripled).

You will only be informed of Group A's decision after the end of the study.

Figure A.4: Overview of instructions (Group B)



Overview

Group B's (your) decision

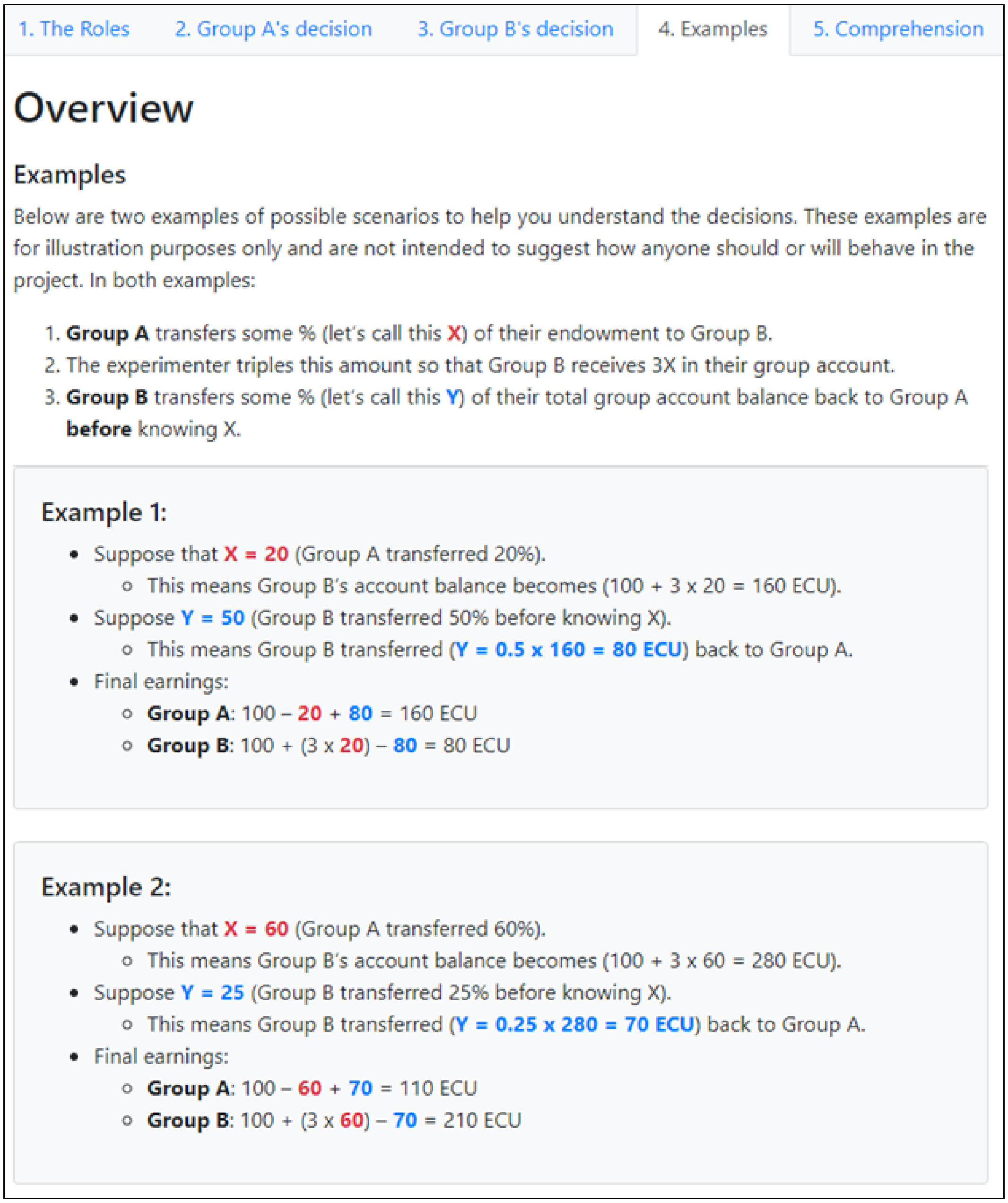
After Group A has made their decision, the amount of ECUs in your group's account is equal to **100 ECU plus 3 times the transfer made by Group A**.

Before you are informed of Group A's transfer, **your group** will be given the chance to have a group discussion and decide what share of the total ECUs in your group's account to transfer back to Group A.

This means you will **not** know the precise balance in your group's account when making this decision. Any amount you transfer will be received directly by Group A (i.e., the transfer will not be tripled).

You will only be informed of Group A's decision after the end of the study.

Figure A.5: Overview of instructions (Group B)



Overview

Examples

Below are two examples of possible scenarios to help you understand the decisions. These examples are for illustration purposes only and are not intended to suggest how anyone should or will behave in the project. In both examples:

1. **Group A** transfers some % (let's call this **X**) of their endowment to Group B.
2. The experimenter triples this amount so that Group B receives 3X in their group account.
3. **Group B** transfers some % (let's call this **Y**) of their total group account balance back to Group A **before** knowing X.

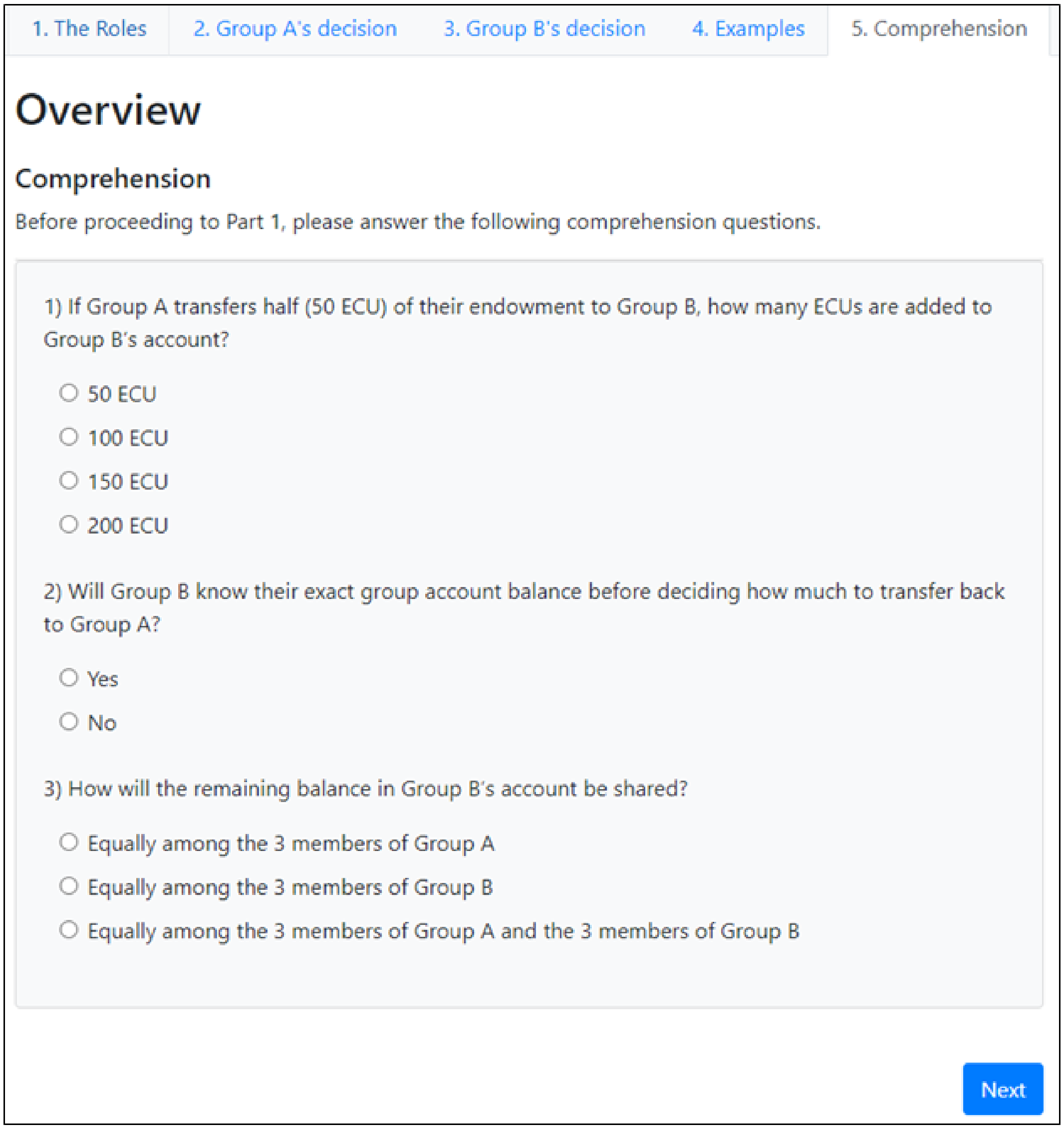
Example 1:

- Suppose that **X = 20** (Group A transferred 20%).
  - This means Group B's account balance becomes (100 + 3 x 20 = 160 ECU).
- Suppose **Y = 50** (Group B transferred 50% before knowing X).
  - This means Group B transferred (**Y = 0.5 x 160 = 80 ECU**) back to Group A.
- Final earnings:
  - **Group A**: 100 – **20** + **80** = 160 ECU
  - **Group B**: 100 + (3 x **20**) – **80** = 80 ECU

Example 2:

- Suppose that **X = 60** (Group A transferred 60%).
  - This means Group B's account balance becomes (100 + 3 x 60 = 280 ECU).
- Suppose **Y = 25** (Group B transferred 25% before knowing X).
  - This means Group B transferred (**Y = 0.25 x 280 = 70 ECU**) back to Group A.
- Final earnings:
  - **Group A**: 100 – **60** + **70** = 110 ECU
  - **Group B**: 100 + (3 x **60**) – **70** = 210 ECU

Figure A.6: Overview of instructions (Group B)



Overview

Comprehension

Before proceeding to Part 1, please answer the following comprehension questions.

1) If Group A transfers half (50 ECU) of their endowment to Group B, how many ECUs are added to Group B's account?

○ 50 ECU
○ 100 ECU
○ 150 ECU
○ 200 ECU

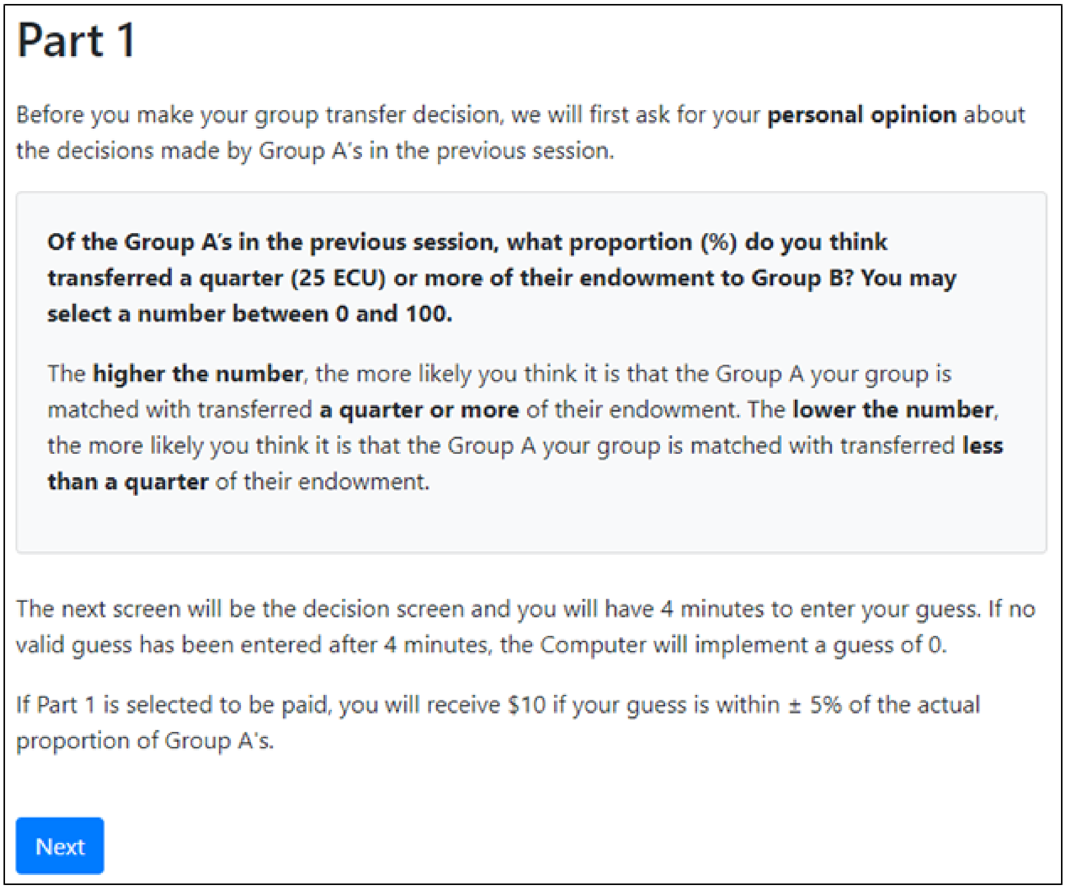
2) Will Group B know their exact group account balance before deciding how much to transfer back to Group A?

○ Yes
○ No

3) How will the remaining balance in Group B's account be shared?

○ Equally among the 3 members of Group A
○ Equally among the 3 members of Group B
○ Equally among the 3 members of Group A and the 3 members of Group B

Figure A.7: Individual prior belief instructions (Group B)



Part 1

Before you make your group transfer decision, we will first ask for your **personal opinion** about the decisions made by Group A's in the previous session.

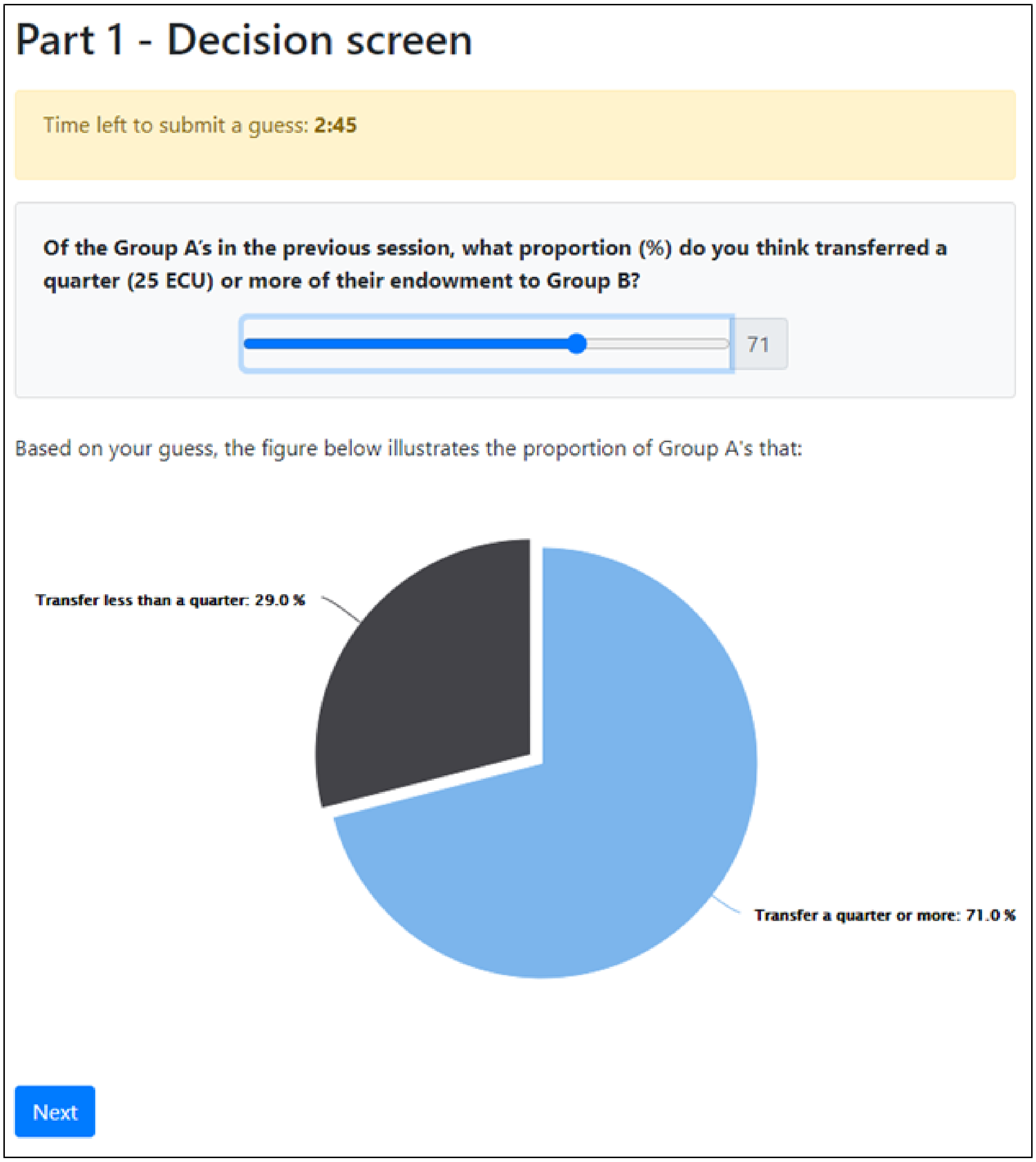
Of the Group A's in the previous session, what proportion (%) do you think transferred a quarter (25 ECU) or more of their endowment to Group B? You may select a number between 0 and 100.

The **higher the number**, the more likely you think it is that the Group A your group is matched with transferred **a quarter or more** of their endowment. The **lower the number**, the more likely you think it is that the Group A your group is matched with transferred **less than a quarter** of their endowment.

The next screen will be the decision screen and you will have 4 minutes to enter your guess. If no valid guess has been entered after 4 minutes, the Computer will implement a guess of 0.

If Part 1 is selected to be paid, you will receive $10 if your guess is within ± 5% of the actual proportion of Group A's.

Figure A.8: Individual prior belief task screen (Group B)



Part 1 - Decision screen

Time left to submit a guess: **2:45**

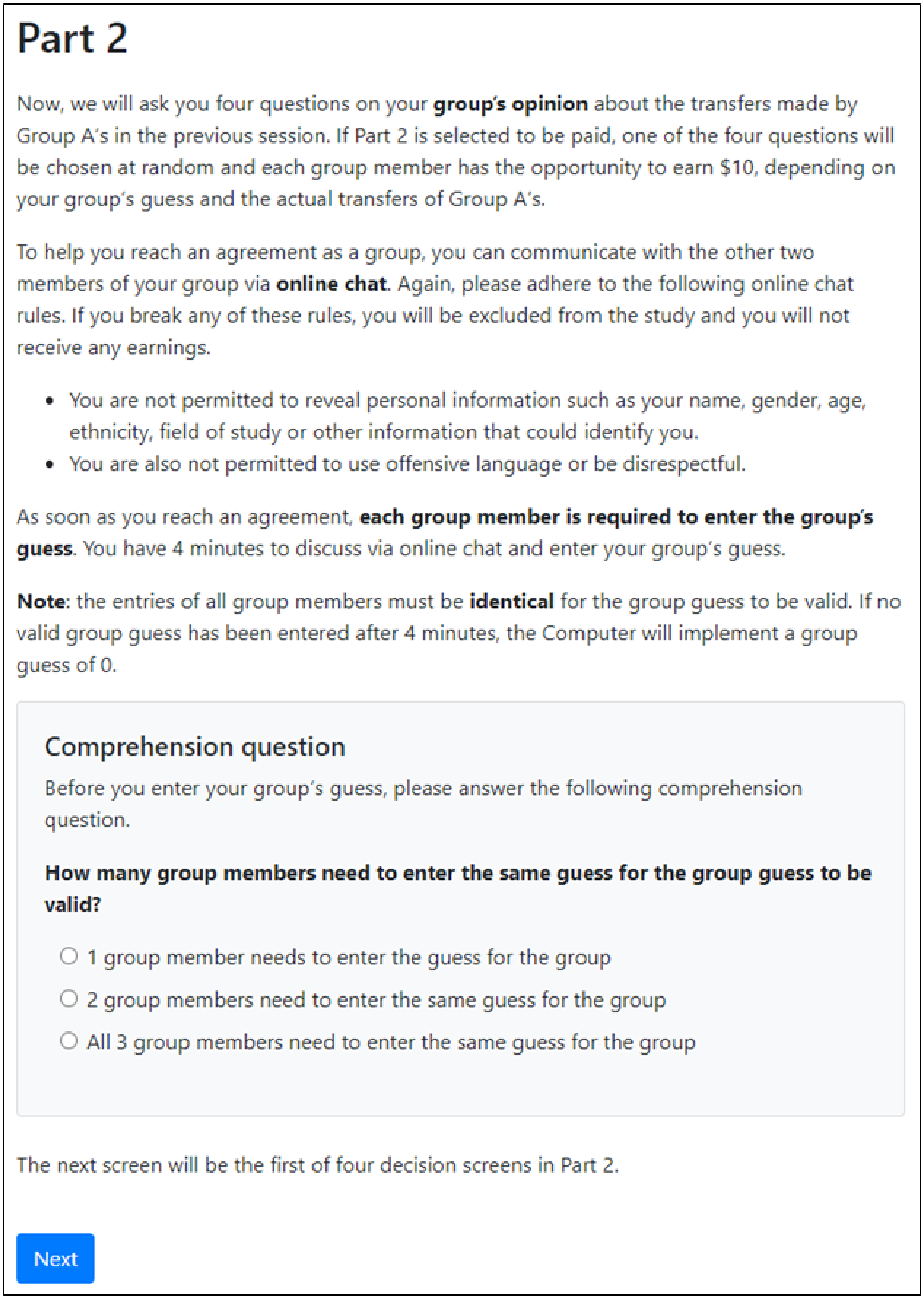
Of the Group A's in the previous session, what proportion (%) do you think transferred a quarter (25 ECU) or more of their endowment to Group B?

71

Based on your guess, the figure below illustrates the proportion of Group A's that:

Transfer less than a quarter: 29.0 %
Transfer a quarter or more: 71.0 %
Next

Figure A.9: Group prior belief instructions (Group B)



Part 2

Now, we will ask you four questions on your **group's opinion** about the transfers made by Group A's in the previous session. If Part 2 is selected to be paid, one of the four questions will be chosen at random and each group member has the opportunity to earn $10, depending on your group's guess and the actual transfers of Group A's.

To help you reach an agreement as a group, you can communicate with the other two members of your group via **online chat**. Again, please adhere to the following online chat rules. If you break any of these rules, you will be excluded from the study and you will not receive any earnings.

- You are not permitted to reveal personal information such as your name, gender, age, ethnicity, field of study or other information that could identify you.
- You are also not permitted to use offensive language or be disrespectful.

As soon as you reach an agreement, **each group member is required to enter the group's guess**. You have 4 minutes to discuss via online chat and enter your group's guess.

Note: the entries of all group members must be **identical** for the group guess to be valid. If no valid group guess has been entered after 4 minutes, the Computer will implement a group guess of 0.

Comprehension question

Before you enter your group's guess, please answer the following comprehension question.

How many group members need to enter the same guess for the group guess to be valid?

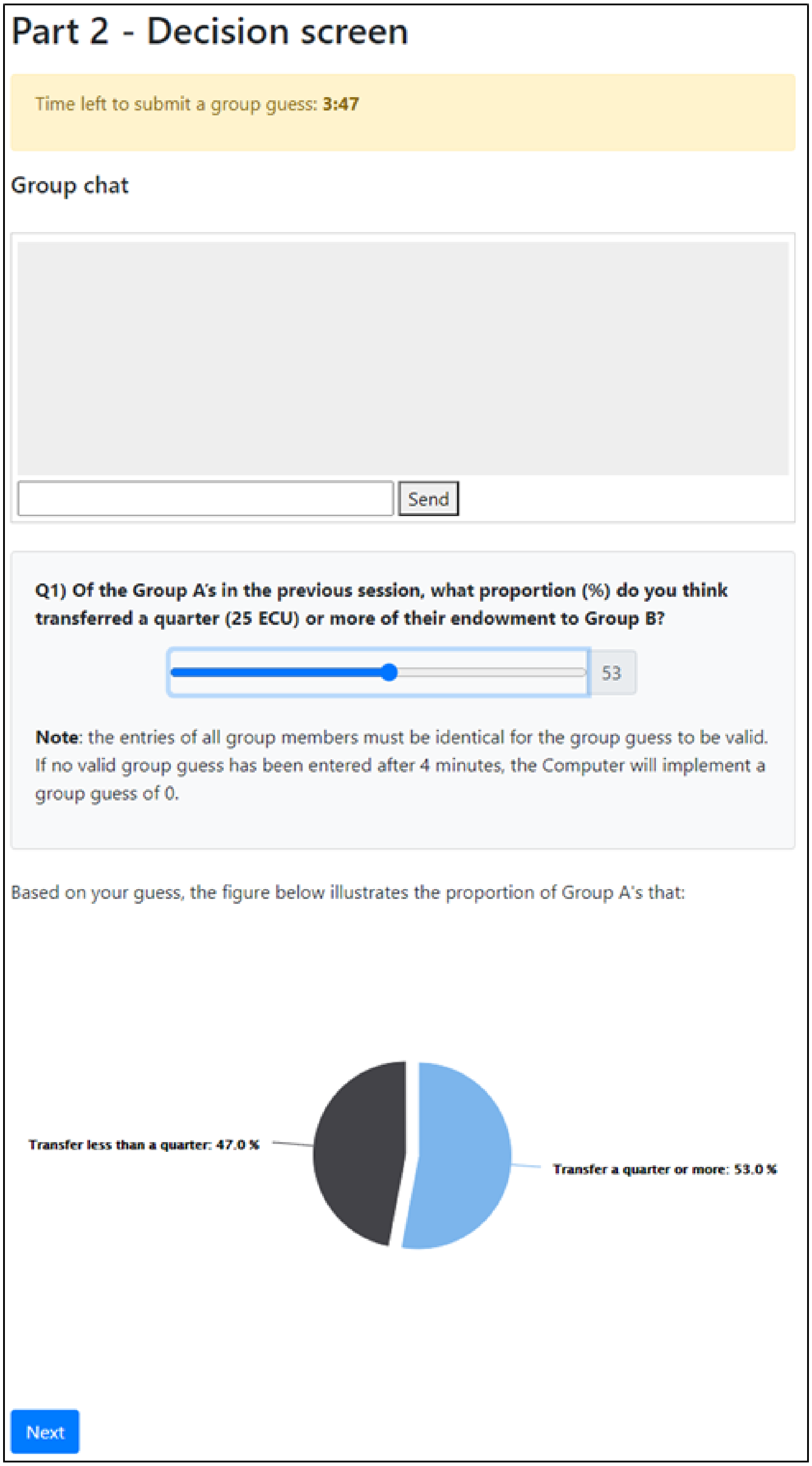
○ 1 group member needs to enter the guess for the group

○ 2 group members need to enter the same guess for the group

○ All 3 group members need to enter the same guess for the group

The next screen will be the first of four decision screens in Part 2.

Figure A.10: Group prior belief task screen (Group B)



Part 2 - Decision screen

Time left to submit a group guess: **3:47**

Group chat

Q1) Of the Group A's in the previous session, what proportion (%) do you think transferred a quarter (25 ECU) or more of their endowment to Group B?

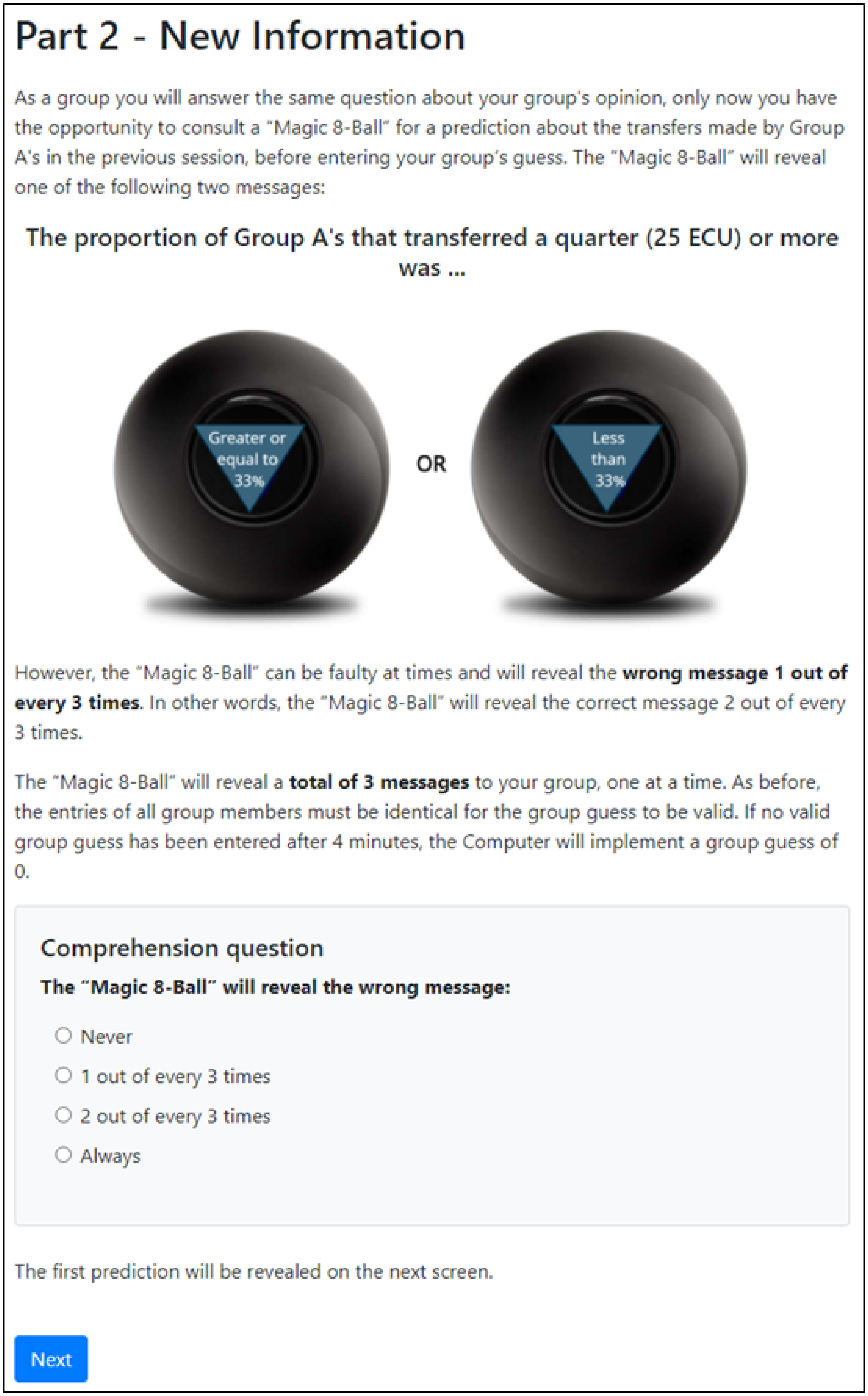
53

Note: the entries of all group members must be identical for the group guess to be valid. If no valid group guess has been entered after 4 minutes, the Computer will implement a group guess of 0.

Based on your guess, the figure below illustrates the proportion of Group A's that:

Transfer less than a quarter: 47.0 %
Transfer a quarter or more: 53.0 %
Next

Figure A.11: Group posterior beliefs instructions (Group B)



Part 2 - New Information

As a group you will answer the same question about your group's opinion, only now you have the opportunity to consult a "Magic 8-Ball" for a prediction about the transfers made by Group A's in the previous session, before entering your group's guess. The "Magic 8-Ball" will reveal one of the following two messages:

The proportion of Group A's that transferred a quarter (25 ECU) or more was ...

However, the "Magic 8-Ball" can be faulty at times and will reveal the **wrong message 1 out of every 3 times**. In other words, the "Magic 8-Ball" will reveal the correct message 2 out of every 3 times.

The "Magic 8-Ball" will reveal a **total of 3 messages** to your group, one at a time. As before, the entries of all group members must be identical for the group guess to be valid. If no valid group guess has been entered after 4 minutes, the Computer will implement a group guess of 0.

Comprehension question

The "Magic 8-Ball" will reveal the wrong message:

○ Never

○ 1 out of every 3 times

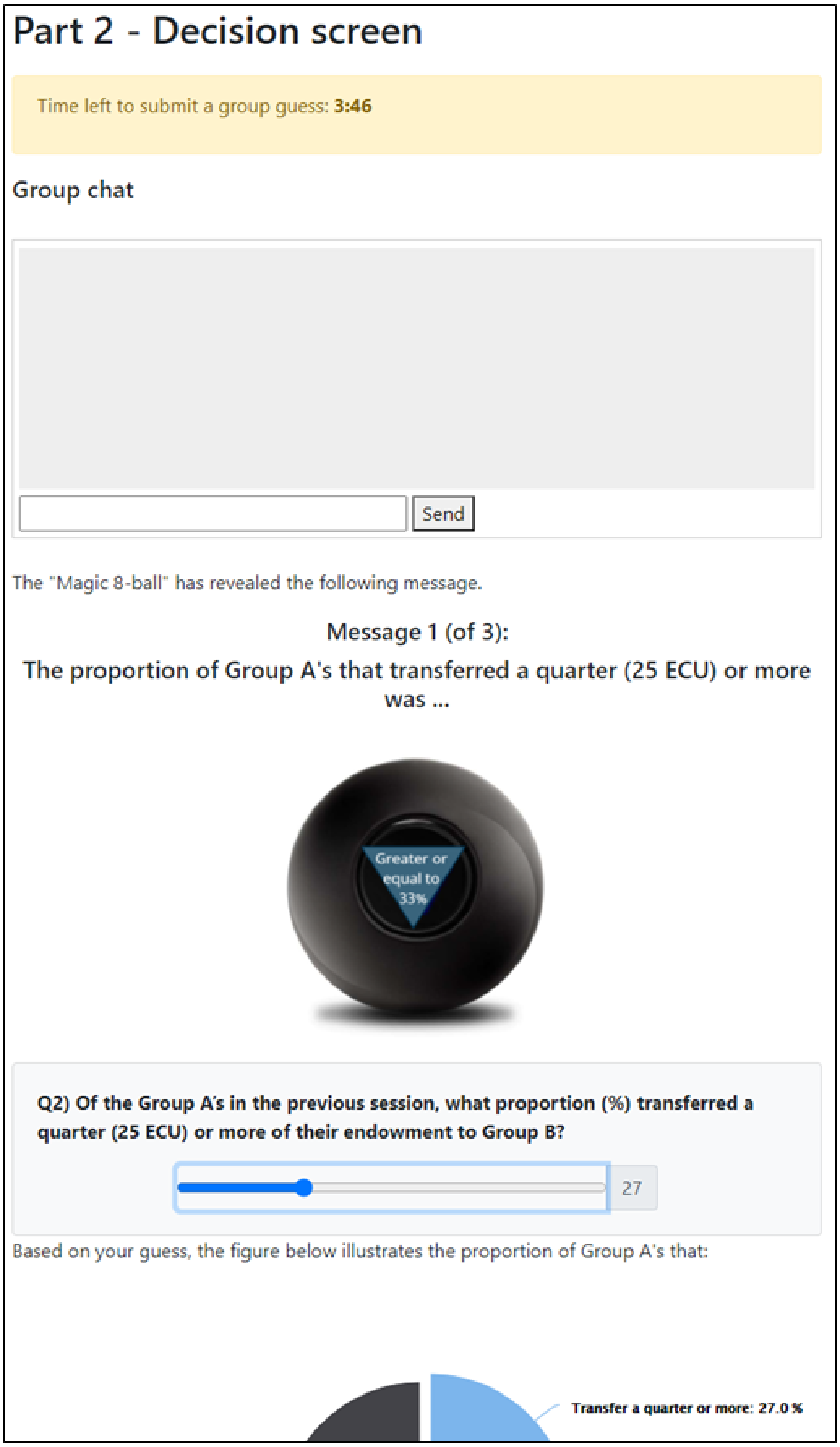
○ 2 out of every 3 times

○ Always

The first prediction will be revealed on the next screen.

Next

Figure A.12: Group posterior beliefs task screen (Group B)



Part 2 - Decision screen

Time left to submit a group guess: **3:46**

Group chat

The "Magic 8-ball" has revealed the following message.

Message 1 (of 3):

The proportion of Group A's that transferred a quarter (25 ECU) or more was ...

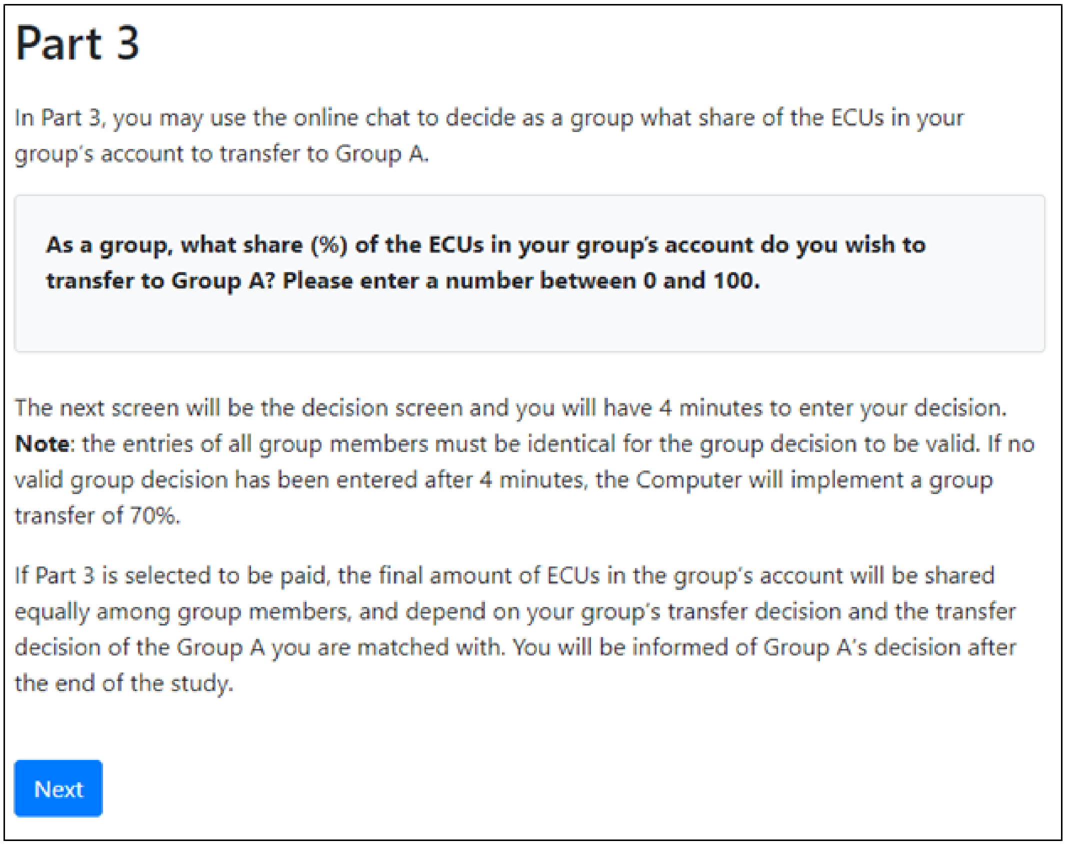
Greater or equal to 33%

Q2) Of the Group A's in the previous session, what proportion (%) transferred a quarter (25 ECU) or more of their endowment to Group B?

27

Based on your guess, the figure below illustrates the proportion of Group A's that:

Figure A.13: Transfer instructions (Group B)



Part 3

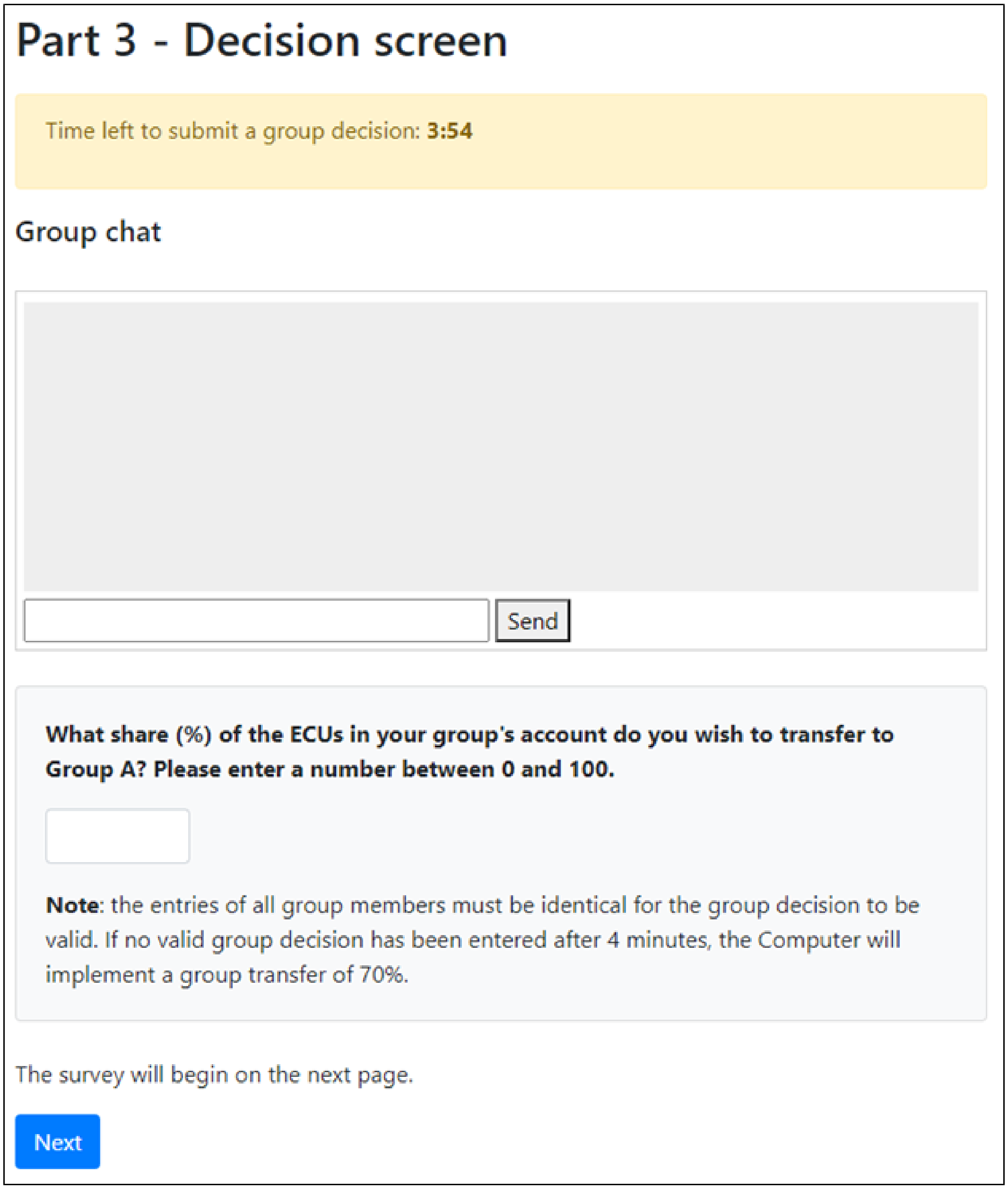
In Part 3, you may use the online chat to decide as a group what share of the ECUs in your group's account to transfer to Group A.

As a group, what share (%) of the ECUs in your group's account do you wish to transfer to Group A? Please enter a number between 0 and 100.

The next screen will be the decision screen and you will have 4 minutes to enter your decision. **Note**: the entries of all group members must be identical for the group decision to be valid. If no valid group decision has been entered after 4 minutes, the Computer will implement a group transfer of 70%.

If Part 3 is selected to be paid, the final amount of ECUs in the group's account will be shared equally among group members, and depend on your group's transfer decision and the transfer decision of the Group A you are matched with. You will be informed of Group A's decision after the end of the study.

Figure A.14: Transfer task screen (Group B)



Part 3 - Decision screen

Time left to submit a group decision: **3:54**

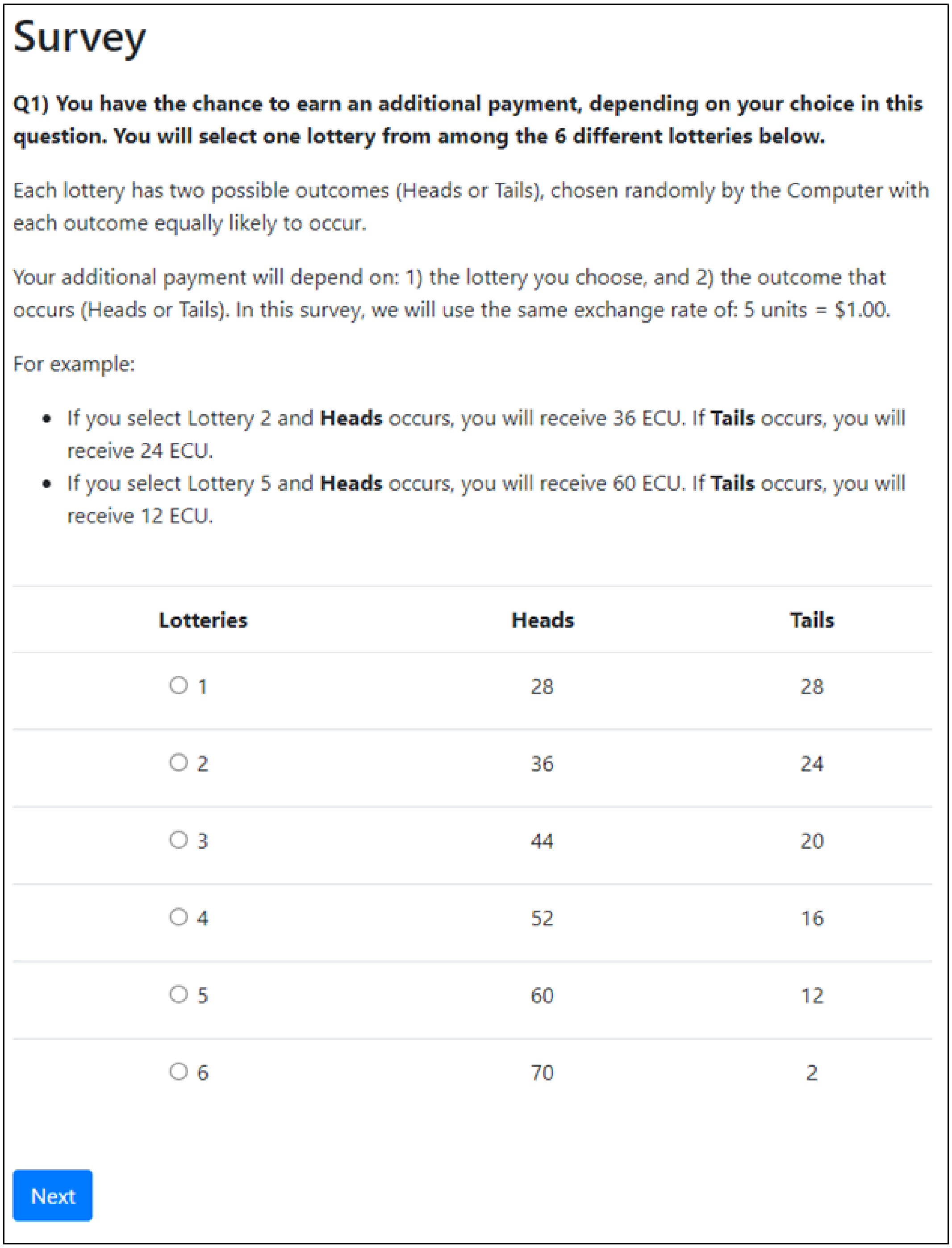
Group chat

What share (%) of the ECUs in your group's account do you wish to transfer to Group A? Please enter a number between 0 and 100.

Note: the entries of all group members must be identical for the group decision to be valid. If no valid group decision has been entered after 4 minutes, the Computer will implement a group transfer of 70%.

The survey will begin on the next page.

Figure A.15: Risk task screen



Survey

Q1) You have the chance to earn an additional payment, depending on your choice in this question. You will select one lottery from among the 6 different lotteries below.

Each lottery has two possible outcomes (Heads or Tails), chosen randomly by the Computer with each outcome equally likely to occur.

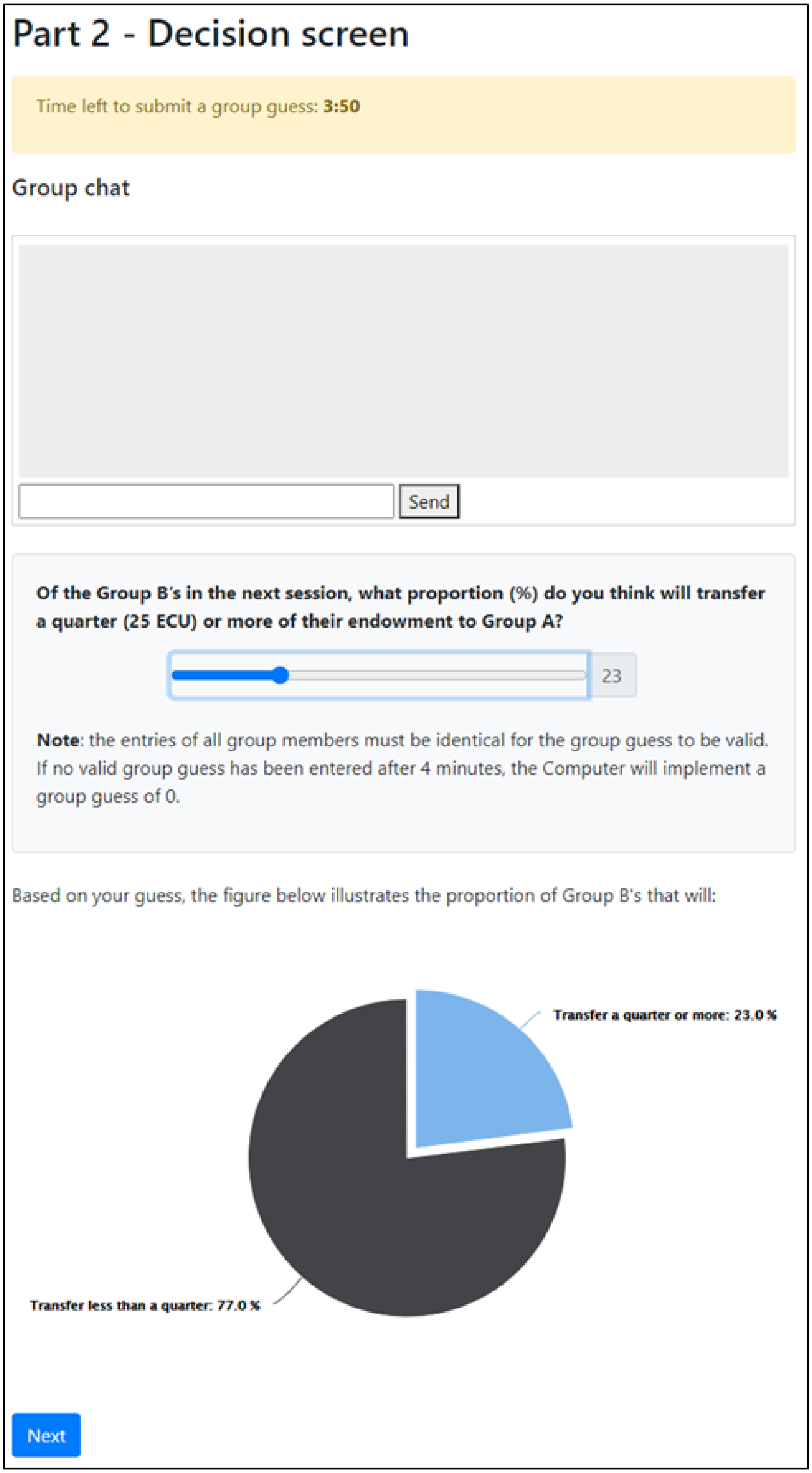
Your additional payment will depend on: 1) the lottery you choose, and 2) the outcome that occurs (Heads or Tails). In this survey, we will use the same exchange rate of: 5 units = $1.00.

For example:

- If you select Lottery 2 and **Heads** occurs, you will receive 36 ECU. If **Tails** occurs, you will receive 24 ECU.
- If you select Lottery 5 and **Heads** occurs, you will receive 60 ECU. If **Tails** occurs, you will receive 12 ECU.

Lotteries	Heads	Tails
○ 1	28	28
○ 2	36	24
○ 3	44	20
○ 4	52	16
○ 5	60	12
○ 6	70	2

Figure A.16: Group prior belief task screen (Group A)



Part 2 - Decision screen

Time left to submit a group guess: **3:50**

Group chat

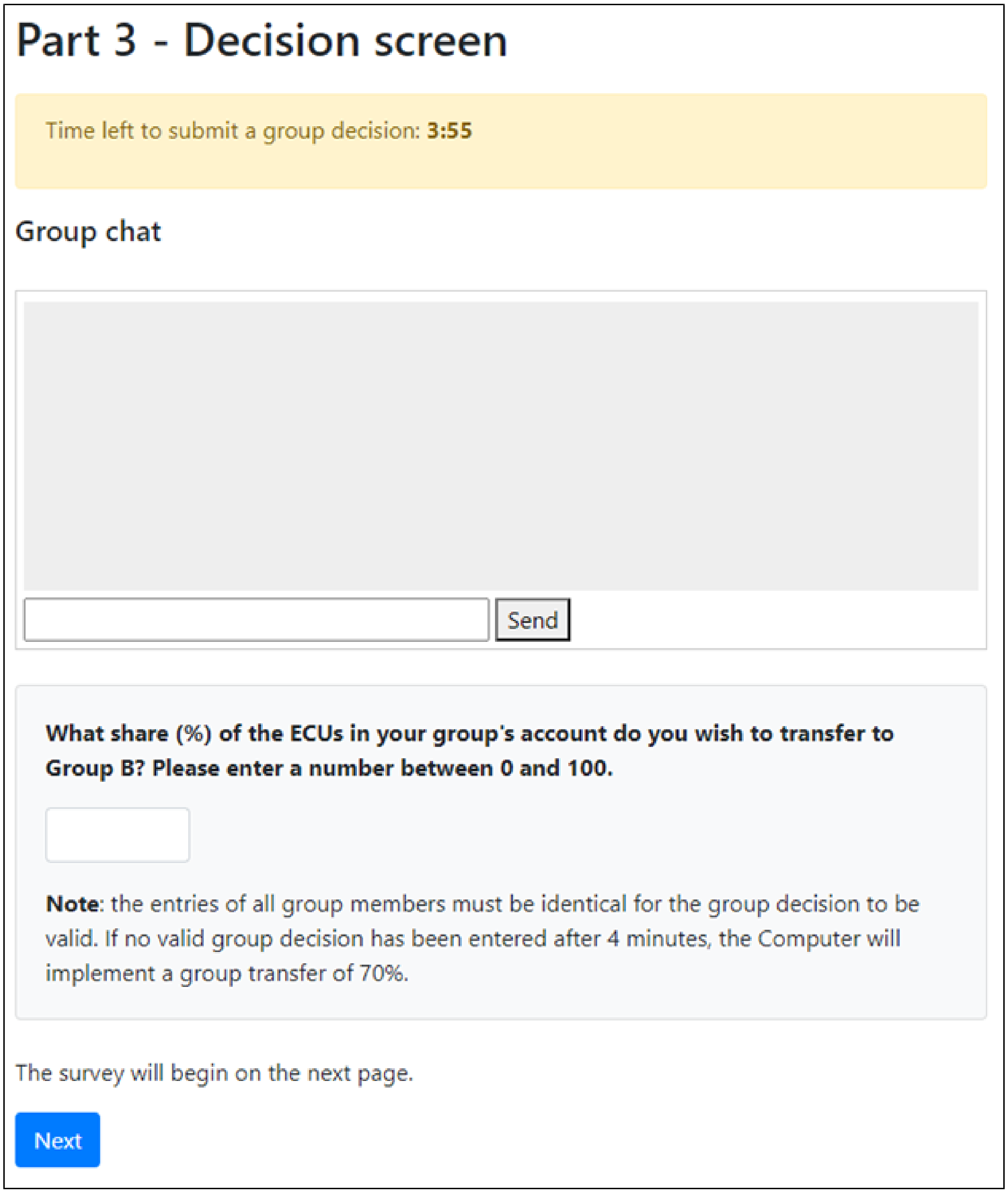
Of the Group B's in the next session, what proportion (%) do you think will transfer a quarter (25 ECU) or more of their endowment to Group A?

Note: the entries of all group members must be identical for the group guess to be valid. If no valid group guess has been entered after 4 minutes, the Computer will implement a group guess of 0.

Based on your guess, the figure below illustrates the proportion of Group B's that will:

Transfer a quarter or more: 23.0 %
Transfer less than a quarter: 77.0 %
Next

Figure A.17: Transfer task screen (Group A)



Part 3 - Decision screen

Time left to submit a group decision: **3:55**

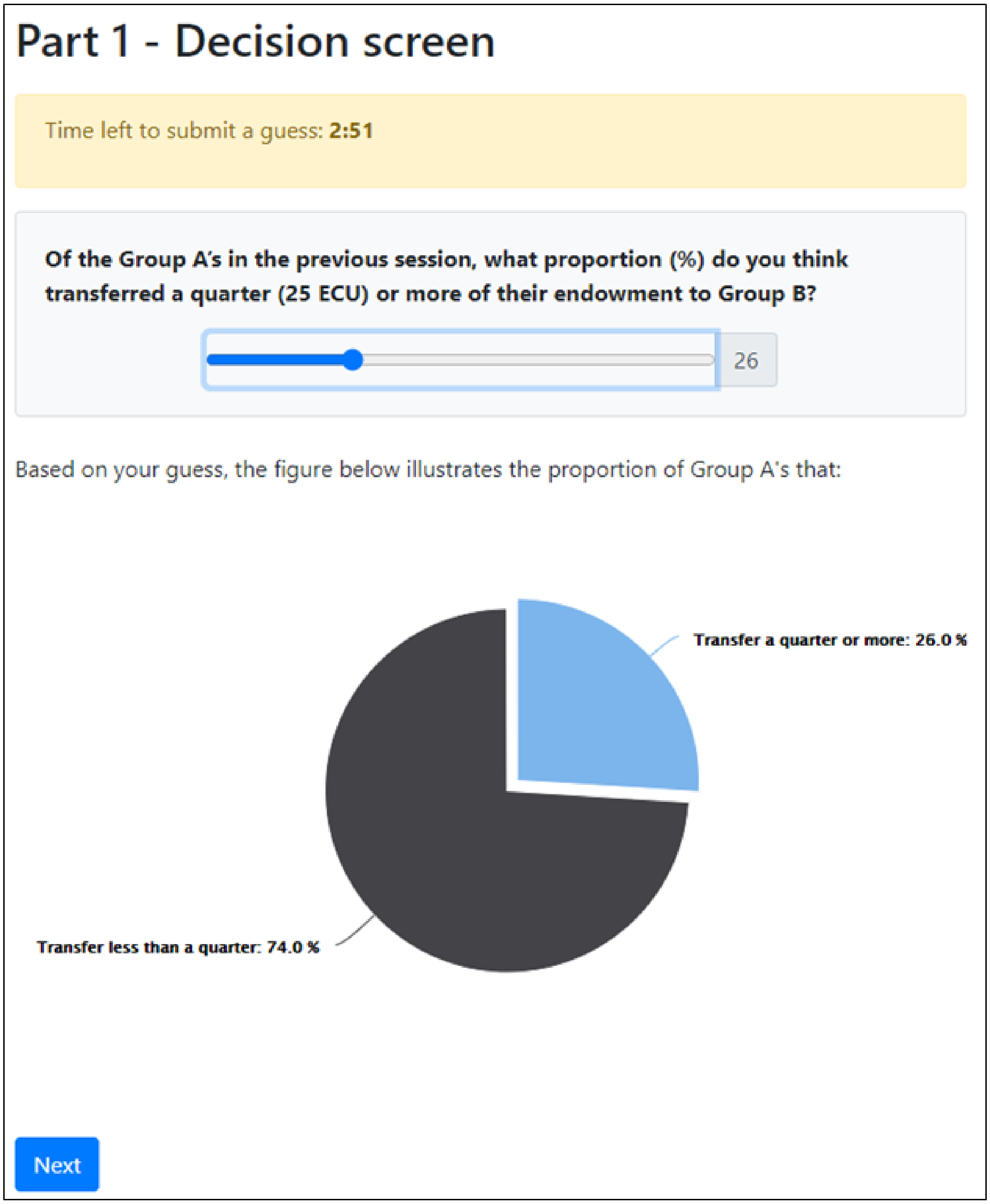
Group chat

What share (%) of the ECUs in your group's account do you wish to transfer to Group B? Please enter a number between 0 and 100.

Note: the entries of all group members must be identical for the group decision to be valid. If no valid group decision has been entered after 4 minutes, the Computer will implement a group transfer of 70%.

The survey will begin on the next page.

Figure A.18: Individual prior belief task screen (Ind B)



Part 1 - Decision screen

Time left to submit a guess: **2:51**

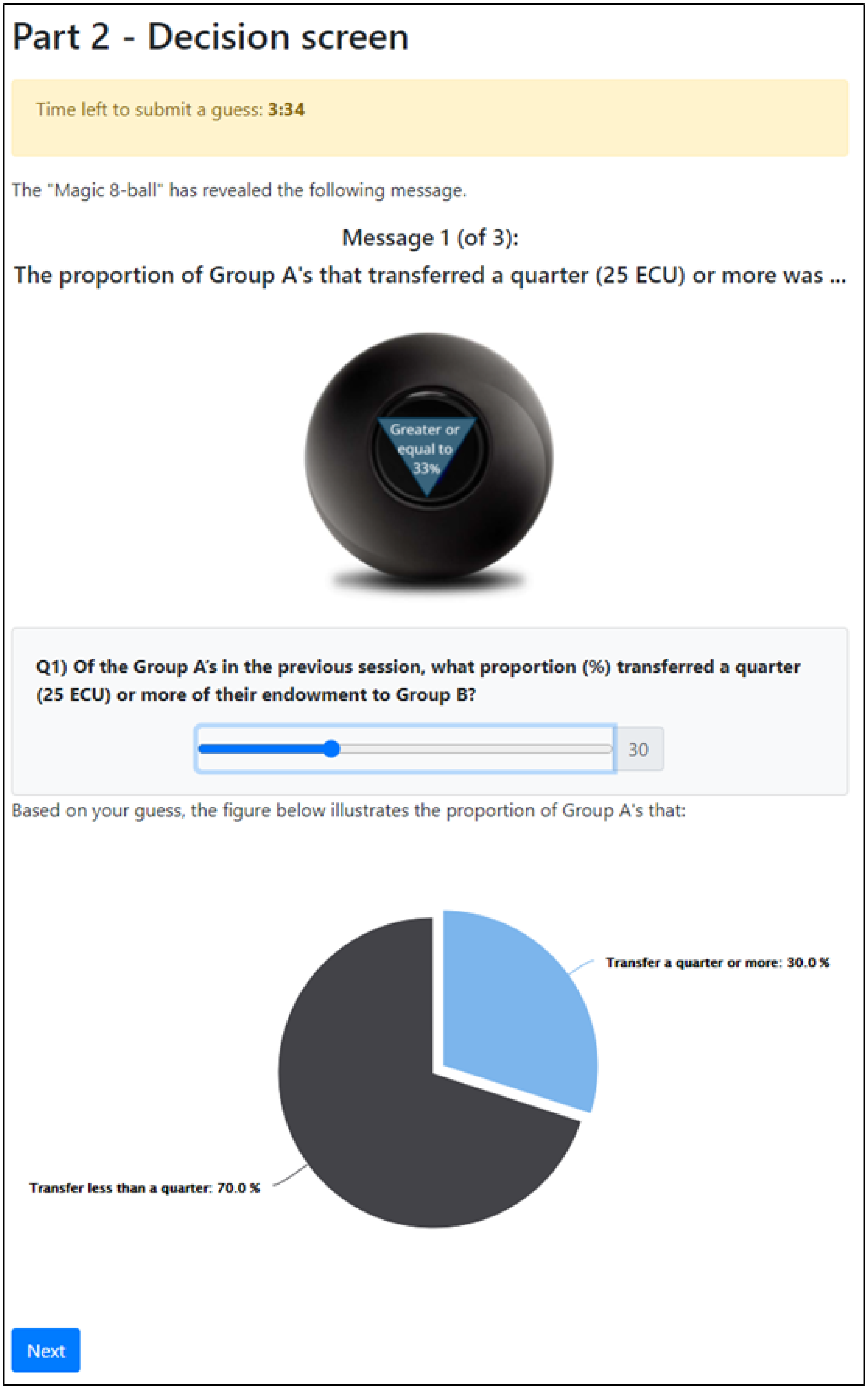
Of the Group A's in the previous session, what proportion (%) do you think transferred a quarter (25 ECU) or more of their endowment to Group B?

26

Based on your guess, the figure below illustrates the proportion of Group A's that:

Transfer a quarter or more: 26.0 %
Transfer less than a quarter: 74.0 %
Next

Figure A.19: Individual posterior belief task screen (Ind B)



Part 2 - Decision screen

Time left to submit a guess: **3:34**

The "Magic 8-ball" has revealed the following message.

Message 1 (of 3):

The proportion of Group A's that transferred a quarter (25 ECU) or more was ...

Greater or equal to 33%

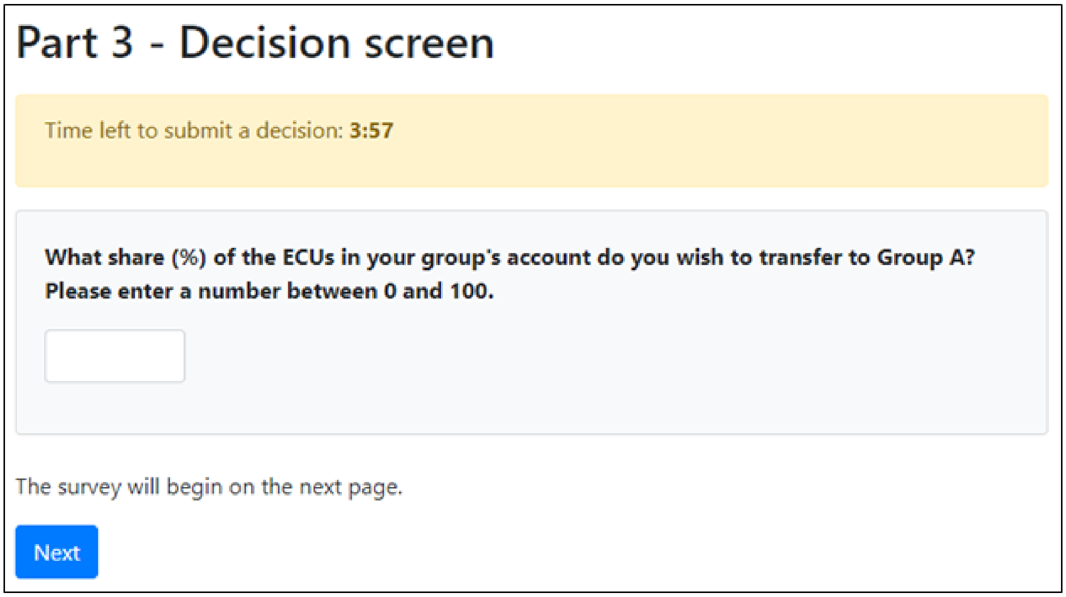
Q1) Of the Group A's in the previous session, what proportion (%) transferred a quarter (25 ECU) or more of their endowment to Group B?

30

Based on your guess, the figure below illustrates the proportion of Group A's that:

Transfer a quarter or more: 30.0 %
Transfer less than a quarter: 70.0 %
Next

Figure A.20: Transfer task screen (Ind B)



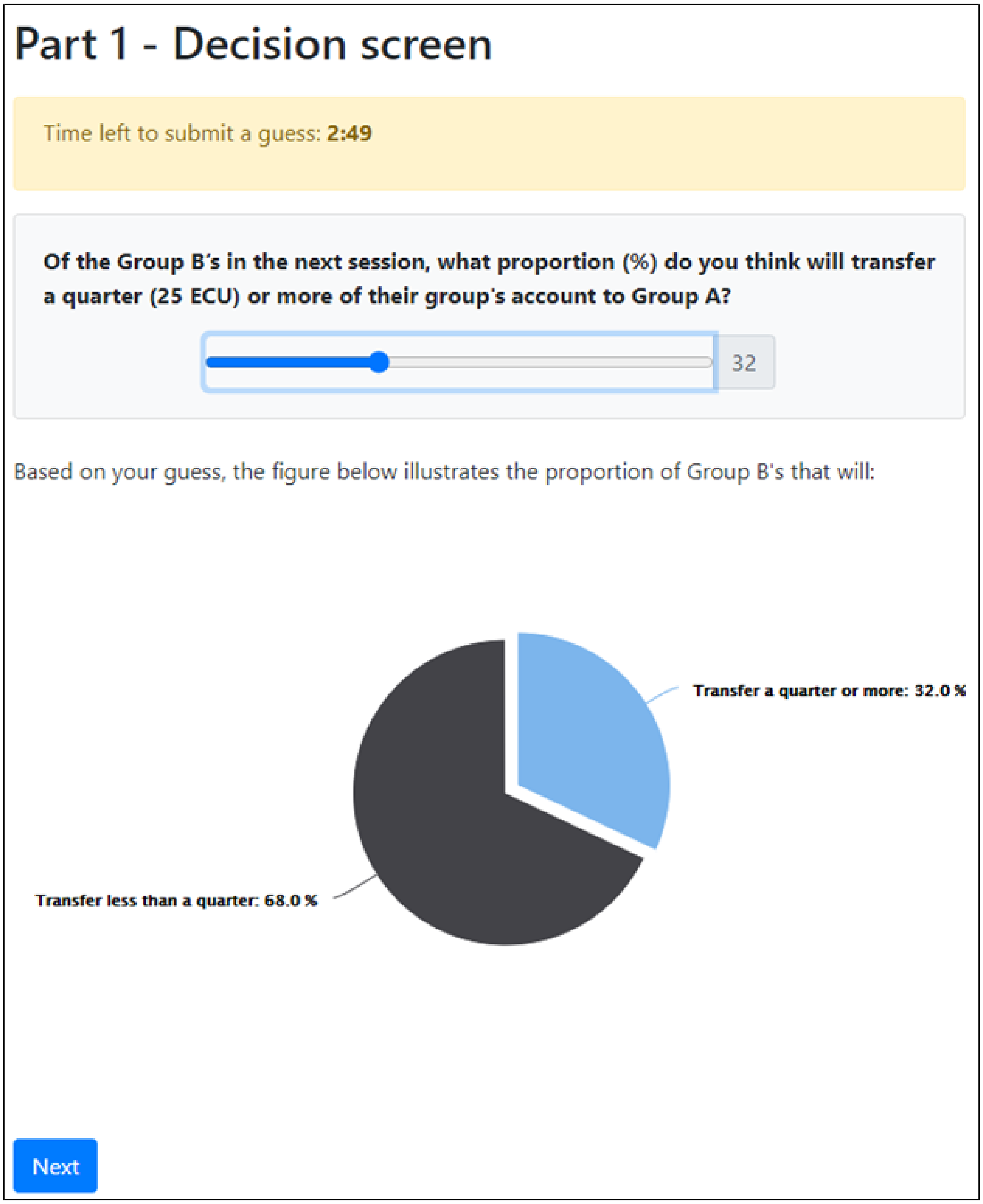
Part 3 - Decision screen

Time left to submit a decision: **3:57**

What share (%) of the ECUs in your group's account do you wish to transfer to Group A? Please enter a number between 0 and 100.

The survey will begin on the next page.

Figure A.21: Individual prior belief task screen (Ind A)



Part 1 - Decision screen

Time left to submit a guess: **2:49**

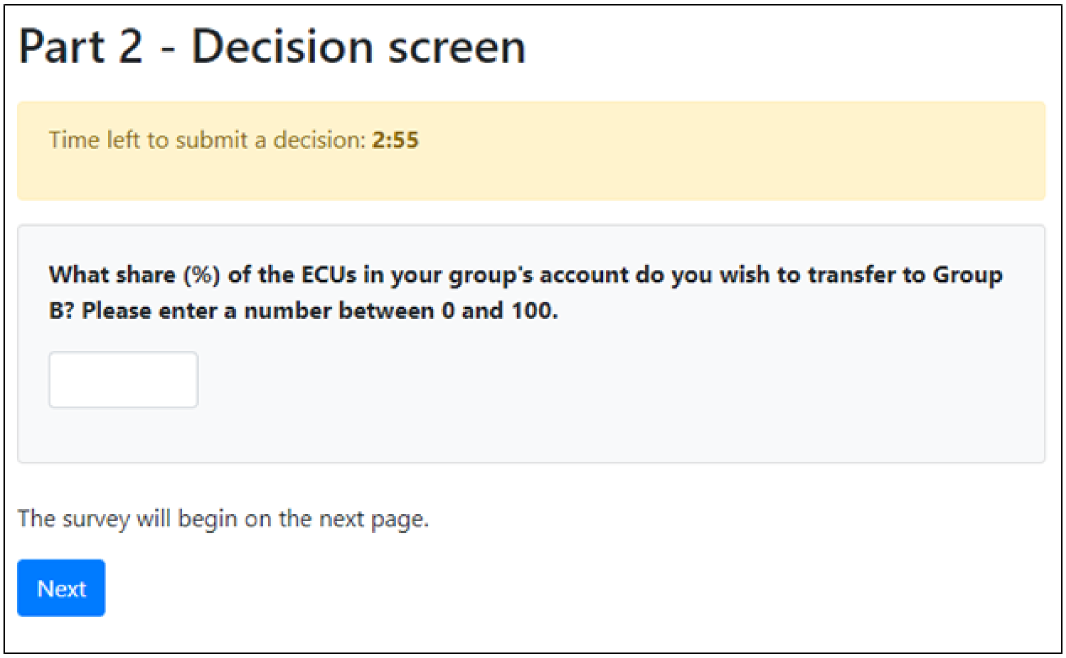
Of the Group B's in the next session, what proportion (%) do you think will transfer a quarter (25 ECU) or more of their group's account to Group A?

32

Based on your guess, the figure below illustrates the proportion of Group B's that will:

Transfer a quarter or more: 32.0 %
Transfer less than a quarter: 68.0 %
Next

Figure A.22: Transfer task screen (Ind A)



Part 2 - Decision screen

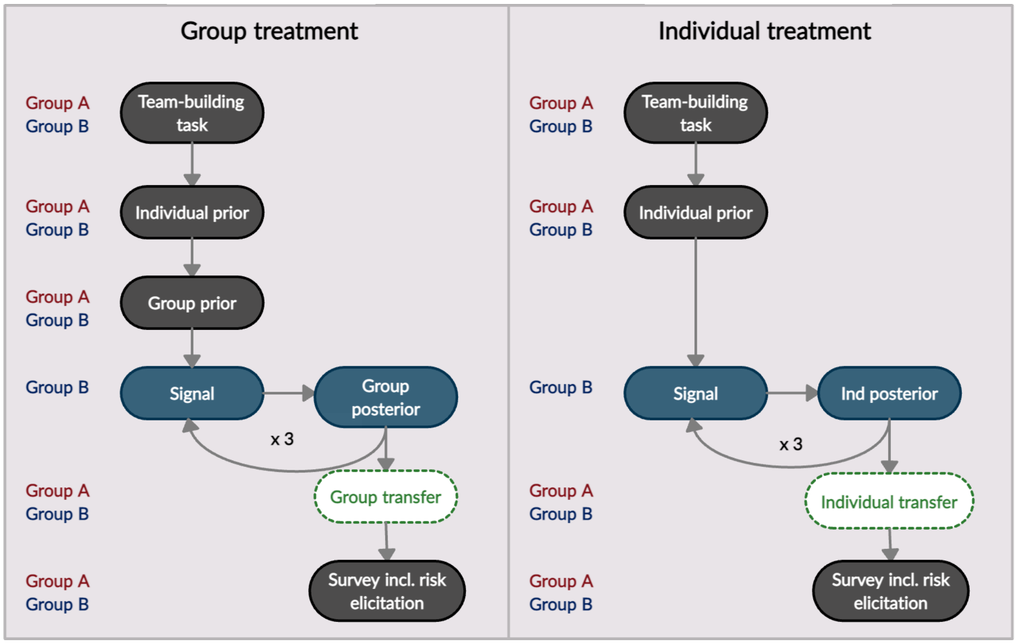
Time left to submit a decision: **2:55**

What share (%) of the ECUs in your group's account do you wish to transfer to Group B? Please enter a number between 0 and 100.

The survey will begin on the next page.

B Experimental timeline

Figure B.1: Timeline



Notes: Subjects in Group B report a total of three posterior beliefs after observing each signal. The dotted lines denote the additional transfer decision made by subjects in *Ind* and *Group* but not in the No-Transfer treatments.

C Confidence in beliefs

C.1 Confidence in prior beliefs

Table C.1: Prior beliefs by confidence in *Ind* and *Group*

	(1)
Confidence	1.32^{**}
	(0.60)
Group	0.11
	(0.58)
Constant	27.83^{***}
	(3.36)
R^2	0.01
Adj. R^2	0.01
Num. obs.	780

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the reported confidence level in prior beliefs. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

C.2 Confidence in posterior beliefs

Table C.2: Posterior beliefs and confidence

	(1)
Confidence	0.95^{*}
	(0.57)
Group	-3.84
	(3.43)
Signal: Low	-7.65^{***}
	(1.65)
Constant	37.15^{***}
	(3.77)
R^2	0.06
Adj. R^2	0.06
Num. obs.	3051

Notes: Ordinary least squares regression with standard errors in parentheses. The dependent variable is the reported confidence level in posterior beliefs. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

D Posterior beliefs

Table D.1 presents average posterior beliefs for individuals and groups, based on the type of signal received. In both *Ind* and *Group*, posterior beliefs are significantly higher following a high signal than a low signal ($p < 0.01$), showing that at the aggregate level, participants are responsive to signals and updating in the correct direction. Following a low signal, individual and group posteriors are not significantly different (33.69 vs. 31.73, $p = 0.27$). However, following a high signal, individual beliefs are significantly higher than group beliefs (43.66 vs. 38.70, $p = 0.04$), suggesting a potential role of communication in the discounting of high signals by groups.

Table D.1: Posterior beliefs by signal type

	Ind	*Group*	*Ind-TP*	*Group-TP*
Signal = Low	33.69	31.73	33.04	33.29
	(2.30)	(1.82)	(2.07)	(2.07)
Signal = High	43.66	38.70	43.62	42.99
	(1.58)	(1.45)	(1.66)	(1.91)
All signals	40.89	36.01	40.44	38.14
	(1.33)	(1.15)	(2.58)	(2.51)

Notes: Standard errors in parentheses. We exclude 3 groups in the Transfer treatments and 4 groups in the Third Party treatments due to a failure to reach an agreement within the allocated time.

Comparing average posterior beliefs in *Ind-TP* and *Group-TP*, posteriors are again significantly lower following a low signal than a high signal ($p < 0.01$). Conditional on receiving a low signal, individuals and groups in the Third party treatments do not differ in their posterior beliefs (33.04 vs. 33.29, $p = 0.61$). We similarly do not observe a difference in posteriors following a high signal (43.62 vs. 42.99, $p = 0.99$). Comparing across the Transfer and Third party treatments, beliefs in *Group-TP* are higher than those in *Group*, with marginal significance (38.14 vs. 36.01, $p = 0.10$).

E Belief updating excluding updates in the wrong direction

Table E.1: Belief updating in *Ind* and *Group*

	Ind (1)	*Group* (2)	Pooled (3)
δ^{prior}	0.53***	0.25***	0.53***
	(0.17)	(0.10)	(0.17)
β^{high}	0.20	−0.33	0.20
	(0.29)	(0.22)	(0.29)
β^{low}	0.49*	0.91***	0.49*
	(0.27)	(0.17)	(0.27)
δ^{prior} x *Group*			0.12
			(0.18)
β^{high} x *Group*			−0.08
			(0.29)
β^{low} x *Group*			−0.06
			(0.56)
$H_0 : \delta^{prior} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{high} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p = 0.06$	$p = 0.62$	$p = 0.77$
$H_0 : \beta^{high} = \beta^{low}$	$p = 0.58$	$p < 0.01$	$p = 0.11$
$H_0 : \beta^{high}$ x *Group* $= \beta^{low}$ x *Group*	-	-	$p = 0.16$
R^2	0.44	0.51	0.47
Adj. R^2	0.44	0.50	0.46
Num. obs.	255	228	483

Notes: Ordinary least squares regression with standard errors clustered at the individual level (in *Ind*) and at the group level (in *Group*) in parentheses. The dependent variable is the belief about the proportion of groups that send a quarter or more of their account. We use individual prior beliefs for *Ind* and group prior beliefs for *Group*. Updates in the wrong direction are excluded. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table E.2: Belief updating in ***Ind-NT*** and ***Group-NT***

	Ind-NT (1)	*Group-NT* (2)	Pooled (3)
δ^{prior}	0.30^{**}	0.42^{***}	0.30^{**}
	(0.12)	(0.13)	(0.12)
β^{high}	-0.14	-0.22	-0.14
	(0.22)	(0.20)	(0.22)
β^{low}	0.85^{*}	0.80^{***}	0.85^{*}
	(0.50)	(0.24)	(0.50)
δ^{prior} x *Group*			0.12
			(0.18)
β^{high} x *Group*			-0.08
			(0.29)
β^{low} x *Group*			-0.06
			(0.56)
$H_0 : \delta^{prior} = 1$	$p = 0.01$	$p < 0.01$	$p = 0.01$
$H_0 : \beta^{high} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p = 0.77$	$p = 0.40$	$p = 0.77$
$H_0 : \beta^{high} = \beta^{low}$	$p = 0.11$	$p = 0.01$	$p = 0.11$
$H_0 : \beta^{high}$ x $Group = \beta^{low}$ x $Group$	-	-	$p = 0.98$
R^2	0.19	0.48	0.29
Adj. R^2	0.18	0.47	0.28
Num. obs.	258	216	474

Notes: Ordinary least squares regression with standard errors clustered at the individual level (in *Ind-NT*) and at the group level (in *Group-NT*) in parentheses. The dependent variable is the belief about the proportion of groups that send a quarter or more of their account. We use individual prior beliefs for *Ind-NT* and group prior beliefs for *Group-NT*. Updates in the wrong direction are excluded. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table E.3: Belief updating for individuals and groups

	Ind (1)	*Group* (2)
δ^{prior}	0.30**	0.42***
	(0.12)	(0.13)
β^{high}	−0.14	−0.22
	(0.22)	(0.20)
β^{low}	0.85*	0.80***
	(0.50)	(0.24)
δ^{prior} x Transfer	0.23	−0.19
	(0.21)	(0.16)
β^{high} Transfer	0.34	−0.09
	(0.36)	(0.28)
β^{low} x Transfer	−0.37	0.06
	(0.57)	(0.29)
$H_0 : \delta^{prior} = 1$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{high} = 1$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p = 0.77$	$p = 0.40$
$H_0 : \beta^{high} = \beta^{low}$	$p = 0.11$	$p = 0.01$
R^2	0.30	0.50
Adj. R^2	0.29	0.49
Num. obs.	513	444

Notes: Ordinary least squares regression with standard errors clustered at the individual level (in *Ind*) and at the group level (in *Group*) in parentheses. The dependent variable is the belief about the proportion of groups that send a quarter or more of their account. We use individual prior beliefs for individuals and group prior beliefs for groups. Updates in the wrong direction are excluded. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table E.4: Belief updating in *Group* by group risk profile

	Group (1)	*Group-NT* (2)	Risk-averse (3)	Risk-tolerant (4)
δ^{prior}	0.26^{***}	0.35^{***}	0.45^{***}	0.40^{***}
	(0.04)	(0.13)	(0.08)	(0.15)
β^{high}	0.01	-0.15	-0.27^{**}	-0.14
	(0.14)	(0.17)	(0.14)	(0.19)
β^{low}	0.54^{**}	0.72^{***}	0.78^{***}	0.75^{***}
	(0.22)	(0.22)	(0.16)	(0.24)
δ^{prior} x Risk-averse	-0.14^{**}	0.10		
	(0.07)	(0.15)		
β^{high} x Risk-averse	-0.96^{***}	-0.12		
	(0.21)	(0.22)		
β^{low} x Risk-averse	0.91^{***}	0.06		
	(0.27)	(0.28)		
δ^{prior} x Transfer			-0.35^{***}	-0.13
			(0.09)	(0.15)
β^{high} x Transfer			-0.64^{***}	0.15
			(0.19)	(0.24)
β^{low} x Transfer			0.59^{***}	-0.21
			(0.21)	(0.33)
$H_0 : \delta^{prior} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{high} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p < 0.01$	$p = 0.18$	$p = 0.18$	$p = 0.29$
$H_0 : \beta^{high} = \beta^{low}$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$
R^2	0.56	0.49	0.66	0.27
Adj. R^2	0.55	0.47	0.65	0.24
Num. obs.	228	216	243	201

Notes: Ordinary least squares regression with standard errors clustered at the group level in parentheses. The dependent variable is the group belief about the proportion of groups that send a quarter or more of their account. Updates in the wrong direction are excluded. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table E.5: **Belief updating in *Ind* by individual risk attitudes**

	Ind	*Ind-NT*	Risk-averse	Risk-tolerant
	(1)	(2)	(3)	(4)
δ^{prior}	0.66^{***}	0.42^{***}	0.37^{**}	0.67^{***}
	(0.17)	(0.15)	(0.17)	(0.13)
β^{high}	0.52	0.03	-0.13	-0.33
	(0.40)	(0.28)	(0.29)	(0.29)
β^{low}	0.40	0.89	0.69^{**}	1.07^{***}
	(0.48)	(0.57)	(0.35)	(0.30)
δ^{prior} x Risk-averse	-0.32	-0.49^{*}		
	(0.22)	(0.27)		
β^{high} x Risk-averse	-0.81^{*}	-0.67^{*}		
	(0.44)	(0.38)		
β^{low} x Risk-averse	0.41	-0.45		
	(0.54)	(0.72)		
δ^{prior} x Transfer			-0.13	-0.48^{***}
			(0.23)	(0.14)
β^{high} x Transfer			-0.16	-0.05
			(0.52)	(0.37)
β^{low} x Transfer			0.42	-0.35
			(0.51)	(0.34)
$H_0 : \delta^{prior} = 1$	$p = 0.04$	$p < 0.01$	$p < 0.01$	$p = 0.01$
$H_0 : \beta^{high} = 1$	$p = 0.23$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p = 0.21$	$p = 0.85$	$p = 0.37$	$p = 0.81$
$H_0 : \beta^{high} = \beta^{low}$	$p = 0.87$	$p = 0.23$	$p = 0.13$	$p < 0.01$
R^2	0.50	0.23	0.56	0.45
Adj. R^2	0.49	0.22	0.55	0.44
Num. obs.	246	258	219	225

Notes: Ordinary least squares regression with standard errors clustered at the individual level in parentheses. The dependent variable is the individual belief about the proportion of groups that send a quarter or more of their account. Updates in the wrong direction are excluded. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

F Belief updating with proxy for non-matches in posterior beliefs

Table F.1: Belief updating in *Ind* and *Group* (proxy for posterior beliefs)

	Ind (1)	*Group* (2)	Pooled (3)
δ^{prior}	0.54^{***}	0.26^{***}	0.54^{***}
	(0.06)	(0.04)	(0.15)
β^{high}	0.21^{*}	-0.31^{***}	0.21
	(0.12)	(0.10)	(0.27)
β^{low}	0.42^{***}	0.89^{***}	0.42^{*}
	(0.16)	(0.13)	(0.25)
δ^{prior} x *Group*			-0.30^{*}
			(0.17)
β^{high} x *Group*			-0.50
			(0.33)
β^{low} x *Group*			0.40
			(0.29)
$H_0 : \delta^{prior} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{high} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p < 0.01$	$p = 0.38$	$p = 0.02$
$H_0 : \beta^{high} = \beta^{low}$	$p = 0.35$	$p < 0.01$	$p = 0.67$
$H_0 : \beta^{high}$ x *Group* $= \beta^{low}$ x *Group*	-	-	$p = 0.11$
R^2	0.45	0.48	0.47
Adj. R^2	0.45	0.48	0.46
Num. obs.	270	270	540

Notes: Ordinary least squares regression with standard errors clustered at the individual level (in *Ind*) and at the group level (in *Group*) in parentheses. The dependent variable is the belief about the proportion of groups that send a quarter or more of their account. We use individual prior beliefs for *Ind* and group prior beliefs for *Group*. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table F.2: Belief updating in *Ind-NT* and *Group-NT* (proxy for posterior beliefs)

	Ind-NT	*Group-NT*	Pooled
	(1)	(2)	(3)
δ^{prior}	0.29^{***}	0.20^{***}	0.29^{**}
	(0.07)	(0.06)	(0.12)
β^{high}	-0.17	-0.19	-0.17
	(0.12)	(0.13)	(0.21)
β^{low}	0.87^{***}	0.81^{***}	0.87^{*}
	(0.28)	(0.14)	(0.46)
δ^{prior} x *Group*			-0.09
			(0.18)
β^{high} x *Group*			-0.03
			(0.32)
β^{low} x *Group*			-0.06
			(0.53)
$H_0 : \delta^{prior} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{high} = 1$	$p < 0.01$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p = 0.64$	$p = 0.19$	$p = 0.77$
$H_0 : \beta^{high} = \beta^{low}$	$p < 0.01$	$p < 0.01$	$p = 0.07$
$H_0 : \beta^{high}$ x *Group* $= \beta^{low}$ x *Group*	-	-	$p = 0.97$
R^2	0.19	0.28	0.23
Adj. R^2	0.18	0.28	0.22
Num. obs.	270	270	540

Notes: Ordinary least squares regression with standard errors clustered at the individual level (in *Ind*) and at the group level (in *Group*) in parentheses. The dependent variable is the belief about the proportion of groups that send a quarter or more of their account. We use individual prior beliefs for *Ind* and group prior beliefs for *Group*. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table F.3: Belief updating by individuals and groups (proxy for posterior beliefs)

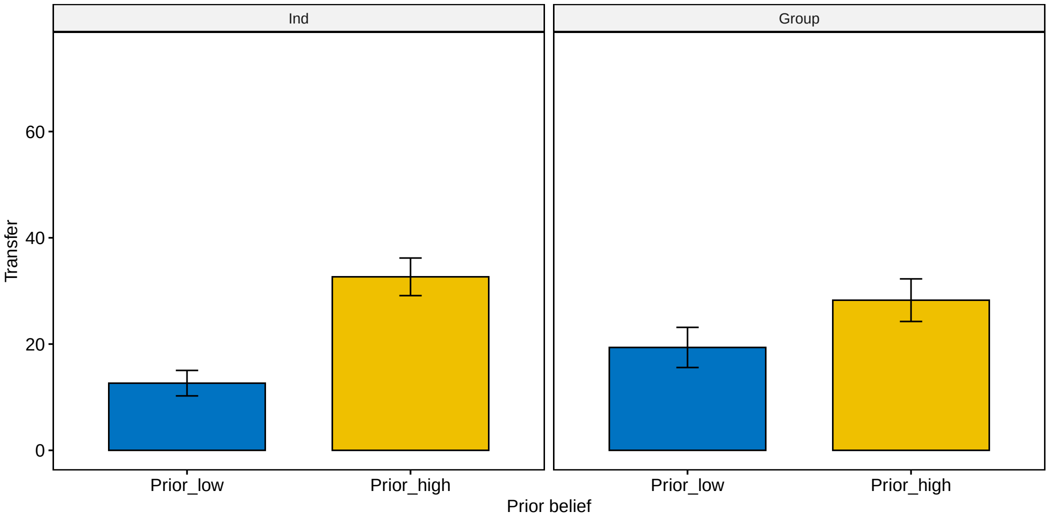
	Ind	*Group*
	(1)	(2)
δ^{prior}	0.29^{**}	0.20
	(0.12)	(0.13)
β^{high}	-0.17	-0.19
	(0.21)	(0.24)
β^{low}	0.87^{*}	0.81^{***}
	(0.46)	(0.27)
δ^{prior} x Transfer	0.25	0.04
	(0.20)	(0.15)
β^{high} x Transfer	0.38	-0.09
	(0.35)	(0.31)
β^{low} x Transfer	-0.44	0.01
	(0.52)	(0.30)
$H_0 : \delta^{prior} = 1$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{high} = 1$	$p < 0.01$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p = 0.77$	$p = 0.48$
$H_0 : \beta^{high} = \beta^{low}$	$p = 0.07$	$p = 0.02$
R^2	0.30	0.37
Adj. R^2	0.30	0.36
Num. obs.	540	540

Notes: Ordinary least squares regression with standard errors clustered at the individual level (in *Ind*) and at the group level (in *Group*) in parentheses. The dependent variable is the belief about the proportion of groups that send a quarter or more of their account. We use individual prior beliefs for *Ind* and group prior beliefs for *Group*. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

G Transfers and beliefs

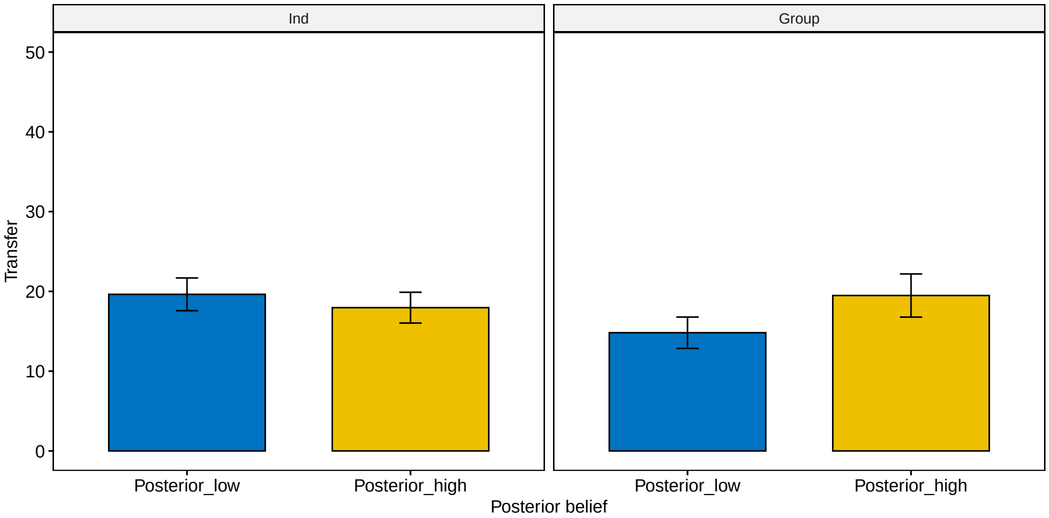
Individuals who have a prior belief that is higher than the median (30%), transfer more than those with a prior that is lower than the median ($p < 0.01$). Similarly, groups with a higher group prior transfer more than groups with a lower group prior belief ($p < 0.01$).

Figure G.1: Transfers by prior belief



Individuals with a posterior that is higher than the median (30%) do not transfer more than those with lower posteriors ($p = 0.52$). Groups with higher posteriors do not transfer more than groups with lower posteriors ($p = 0.14$). Groups with low posteriors transfer less than individuals ($p = 0.02$) but we do not observe this for higher posteriors ($p = 0.68$)

Figure G.2: Transfers by posterior beliefs (Group B)



H Latent dirichlet allocation

Figure H.1: Selecting the optimal number of topics (Group A)

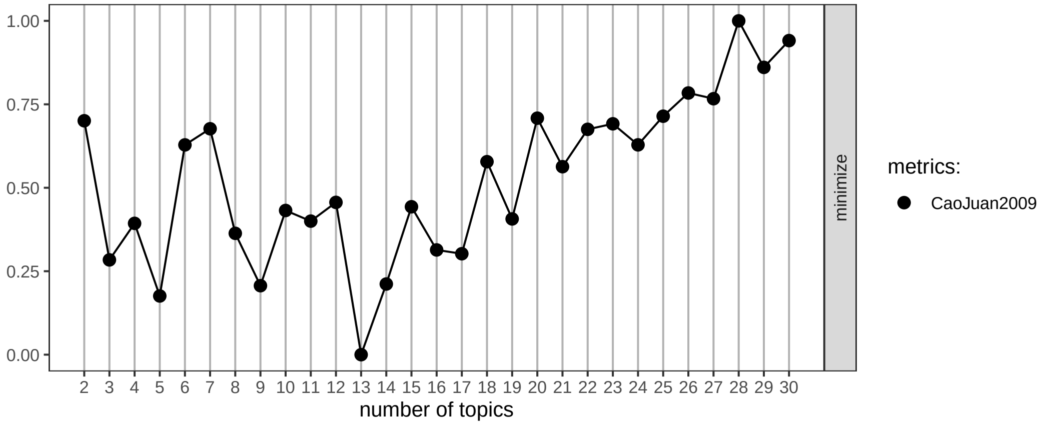


Figure H.2: Topics in the Latent Dirichlet Allocation model (Group A)

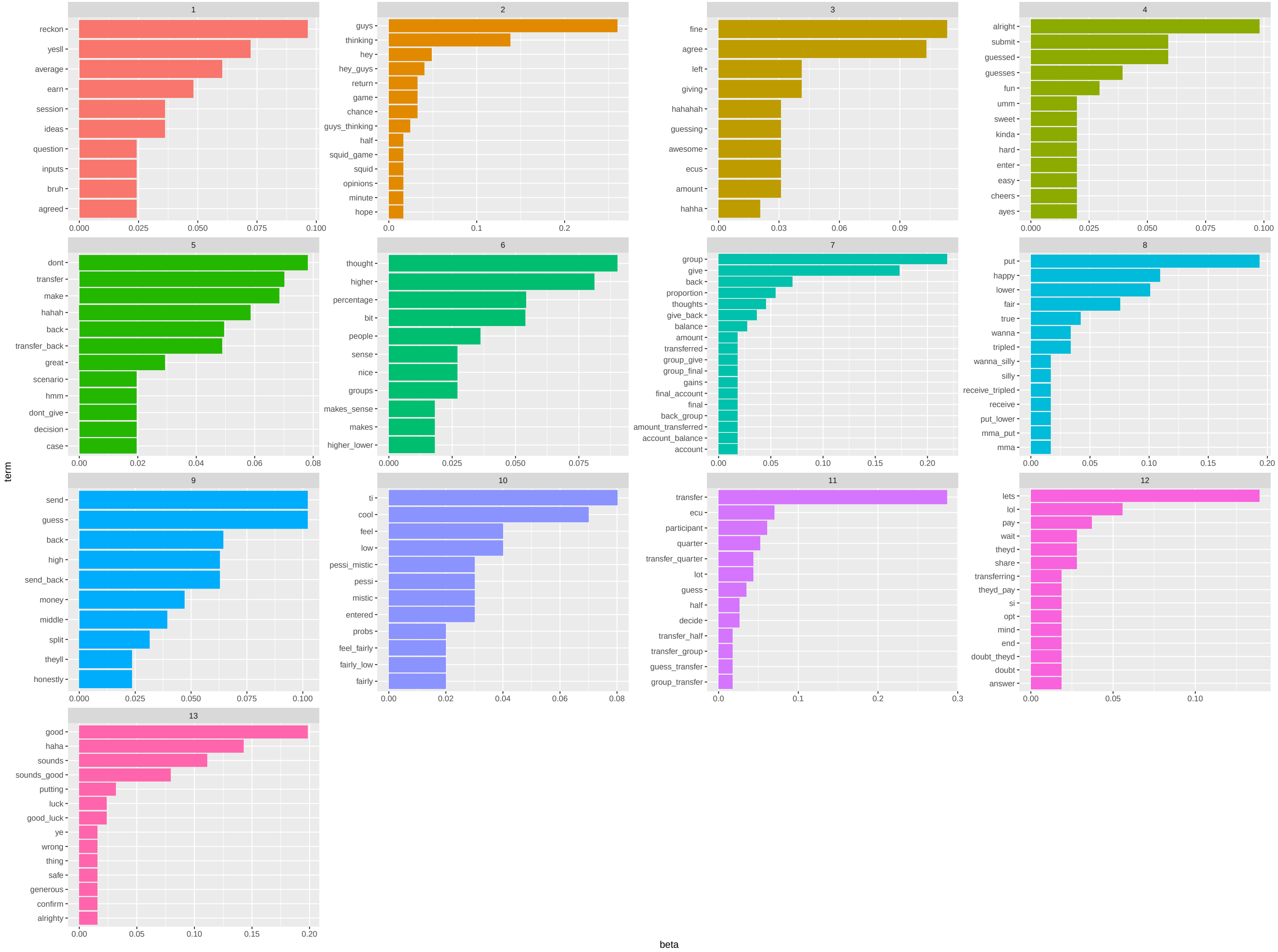
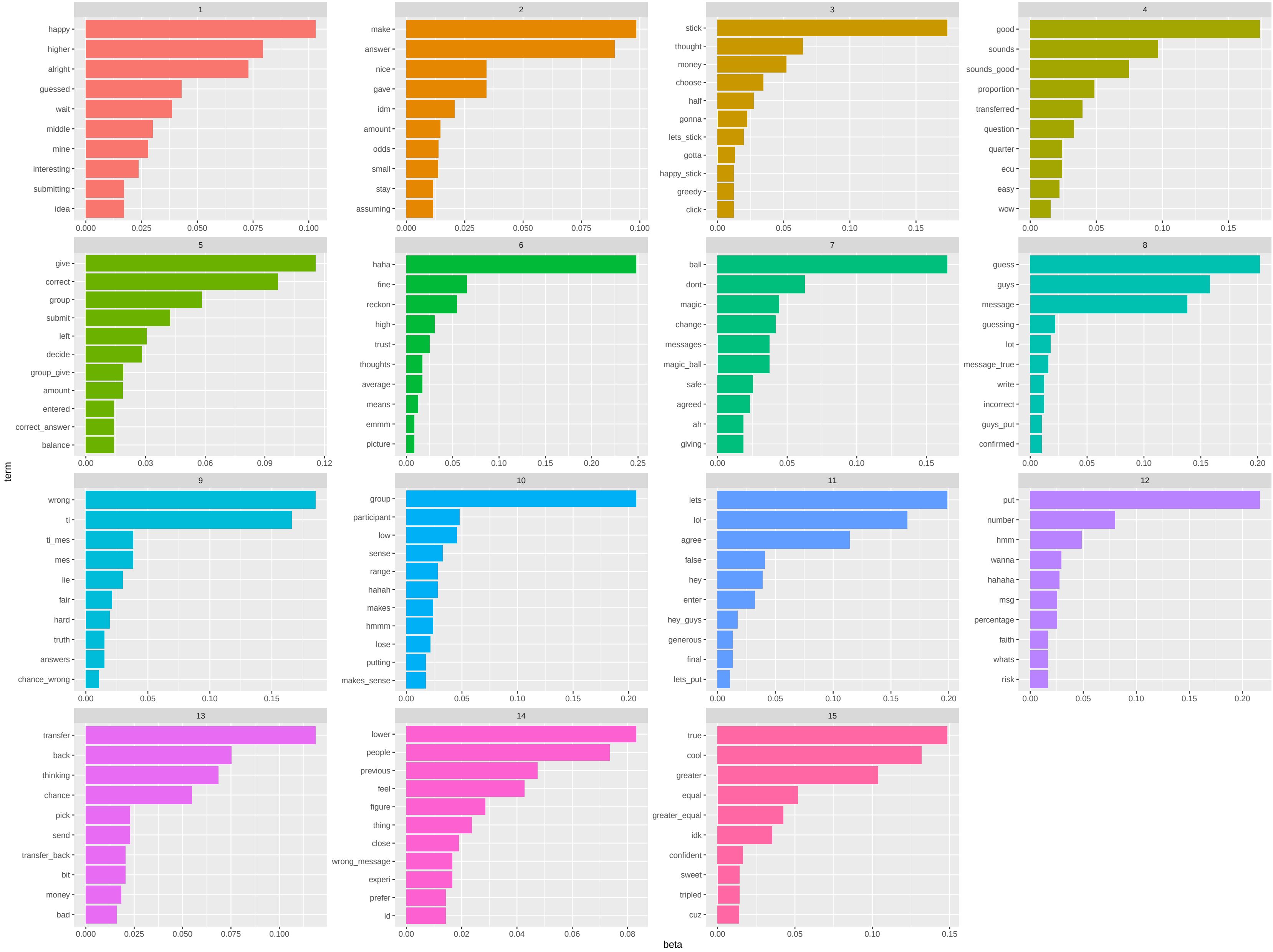


Figure H.3: Topics in the Latent Dirichlet Allocation model (Group B)



I Risk preferences

Figure I.1 depicts the distribution of lottery choices (with 6 choices in total) across all treatments, with lottery 1 being the least risky and 6 being the most risky. Consistent with the literature (e.g., Eckel and Grossman, 2002; Croson and Gneezy, 2009), men are on average more likely to choose a riskier lottery compared to women (3.82 vs. 3.16, $p < 0.01$).

Comparing risk attitudes across the Transfer and No-Transfer treatments (see Figure I.2), we find no significant difference in the distribution of individual risk preferences ($p = 0.76$, Kolmogorov-Smirnov test). On average, subjects in the Transfer treatment have a risk measure of 3.47 while those in the No-Transfer treatment report a risk measure of 3.58 ($p = 0.53$). Similarly, average risk is 3.52 in *Ind* and 3.50 in *Group* and this difference is not significant ($p = 0.88$).

Figure I.1: Individual risk attitudes

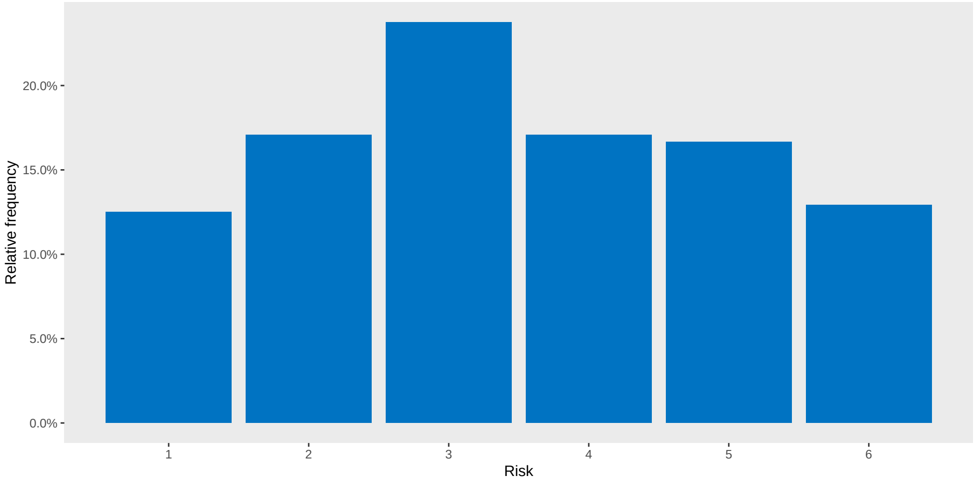
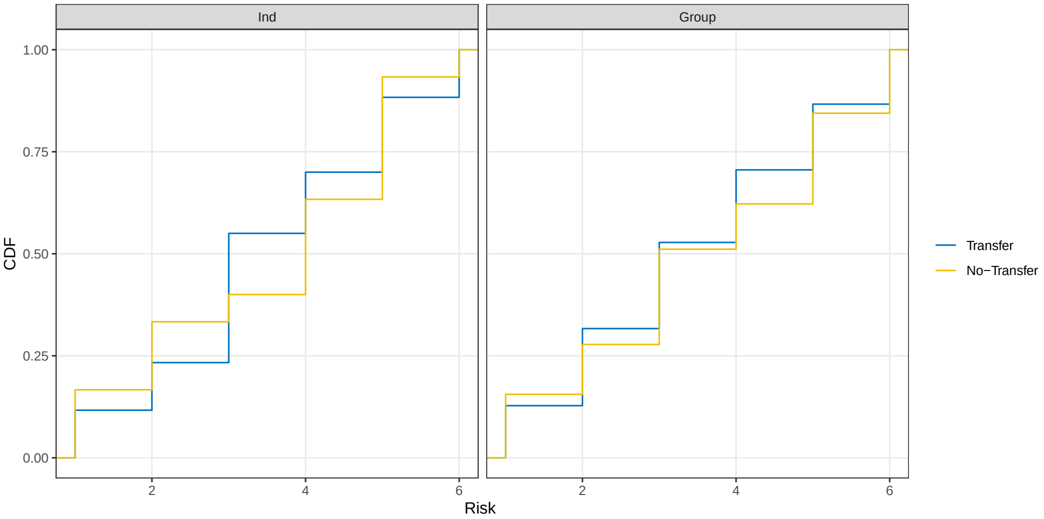


Figure I.2: Individual risk attitudes in *Ind* and *Group*



J Belief updating by the amount transferred in the trust game

Table J.4 presents the regression analysis based on whether individuals and groups transferred a relatively low (high) amount, i.e., less (more) than the median transfer of of 17.70%, which we use as a proxy for social preferences. Individuals who transfer less than the median do not differ from those who transfer more than the median in the weights placed on prior beliefs (column 1, $p = 0.56$), high signals ($p = 0.25$) and low signals ($p = 0.72$). Similarly, we find no significant difference between groups that transfer low and high amounts for prior beliefs (column 2, $p = 0.38$), high signals ($p = 0.28$) and low signals ($p = 0.28$). For those that transfer a lower amount, groups tend to be less responsive to high signals than individuals (column 3, $p = 0.04$), and we find some evidence that low-transfer groups are more asymmetric in belief updating than low-transfer individuals. Given that we do not observe the transfer decisions of third parties, we are unable to directly observe self-serving biases.

Table J.4: Belief updating by transfer

	Ind	*Group*	Low-transfer	High-transfer
	(1)	(2)	(3)	(4)
δ^{prior}	0.67^{***}	0.39^{**}	0.52^{***}	0.66^{***}
	(0.19)	(0.17)	(0.19)	(0.19)
β^{high}	-0.08	-0.09	0.46	-0.01
	(0.25)	(0.32)	(0.41)	(0.26)
β^{low}	0.55^{**}	0.55	0.37	0.44^{*}
	(0.25)	(0.40)	(0.45)	(0.27)
δ^{prior} x Low-transfer	-0.16	-0.17		
	(0.27)	(0.20)		
β^{high} x Low-transfer	0.55	-0.44		
	(0.48)	(0.41)		
β^{low} x Low-transfer	-0.18	0.47		
	(0.51)	(0.44)		
δ^{prior} x Group			-0.32	-0.29
			(0.22)	(0.25)
β^{high} x Group			-0.96^{**}	-0.08
			(0.47)	(0.39)
β^{low} x Group			0.57	0.07
			(0.48)	(0.46)
$H_0 : \delta^{prior} = 1$	$p = 0.09$	$p < 0.01$	$p = 0.01$	$p = 0.07$
$H_0 : \beta^{high} = 1$	$p < 0.01$	$p < 0.01$	$p = 0.19$	$p < 0.01$
$H_0 : \beta^{low} = 1$	$p = 0.07$	$p = 0.26$	$p = 0.16$	$p = 0.04$
$H_0 : \beta^{high} = \beta^{low}$	$p = 0.16$	$p = 0.34$	$p = 0.91$	$p = 0.35$
$H_0 : \beta^{high}$ x *Group* $= \beta^{low}$ x *Group*	-	-	$p = 0.08$	$p = 0.85$
R^2	0.50	0.50	0.48	0.50
Adj. R^2	0.49	0.49	0.47	0.49
Num. obs.	261	249	273	246

Notes: Ordinary least squares regression with standard errors clustered at the individual level in parentheses. The dependent variable is the individual belief about the proportion of groups that send a quarter or more of their account. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Chapter 3: Norm-signalling punishment[*]

Daniele Nosenzo [†], Erte Xiao [‡], Nina Xue [§]

Abstract

The literature on punishment and prosocial behavior has presented conflicting findings. In some settings, punishment crowds out prosocial behavior and backfires; in others, however, it promotes prosociality. We examine whether the punisher's motives can help reconcile these results through a novel experiment in which the agent's outcomes are identical in two environments, but in one punishment is self-serving (i.e., potentially benefits the punisher) while in the other it is other-regarding (i.e., potentially benefits a third party). We find that self-regarding punishment reduces the social stigma of selfish behavior, while other-regarding punishment does not. As a result, self-serving punishment is less effective at encouraging compliance and is more likely to backfire compared to other-regarding punishment. Our findings have implications for the design of punishment mechanisms and highlight the importance of the punisher's motives in the norm-signalling function of punishment.

JEL Classification: C91, C72, D02
Keywords: punishment, norms, stigma, crowd out, experiment

[*] We received helpful comments from Eugen Dimant, Nick Feltovich, Xiaojian Zhao and audiences at the Norms and Behavioral Change Talk and Monash University. This work was supported by the Australian Research Council (DPDP16010274) and Aarhus University Research Foundation (AUFF Starting Grant 36835) and received approval from the Monash University Human Research Ethics Committee (project number 27176).

[†] Aarhus Univeristy, Denmark, Daniele.Nosenzo@econ.au.dk
[‡] Monash University, Australia, Erte.Xiao@monash.edu
[§] Monash University, Australia, Nina.Xue@monash.edu

1 Introduction

Evidence on the effectiveness of punishment in disciplining individual self-interest is mixed. In some settings, punishment appears to effectively restrain self-interest and promote prosocial behavior (e.g., Fehr and Gächter, 2002; Andreoni et al., 2003; Villatoro et al., 2014). However, another line of research shows that punishment can sometimes backfire and crowd out prosocial behavior (e.g., Gneezy and Rustichini, 2000; Fehr and Rockenbach, 2003; Galbiati et al., 2013).[1] The conflicting findings raise the question of why punishment crowds in prosocial behavior in some cases, but crowds out prosociality in others.

Scholars in law and economics have argued that an important function of punishment – which is crucial for its effectiveness – is to communicate information about society's norms and values (e.g., Sunstein, 1996; Posner, 1997; Kahan, 1998; McAdams, 2000; Bénabou and Tirole, 2006; Bénabou and Tirole, 2011). This paper investigates whether the punisher's *motivation* for imposing punishment can affect the message conveyed about underlying social norms, hence altering its effectiveness. In particular, we compare two forms of punishment: (1) punishment that is designed to nudge an agent towards compliance for the punisher's own gain ("self-serving punishment"), and (2) punishment that encourages compliance for the benefit of a third party ("other-regarding punishment"). Based on a simple theoretical framework, our hypothesis is that self-serving punishment transmits a weaker normative message compared to other-regarding punishment, and is hence more likely to trigger a crowding-out of prosocial behavior.

We design a novel principal-agent experimental paradigm to examine whether the same punishment mechanism (from the agent's perspective) can both crowd in and crowd out prosocial behavior, depending on whether punishment is motivated by self-interest, or by a concern for others. A key feature of our design is that our treatments hold the agent's payoffs constant, and only differ in whether the (credible) threat of punishment can be used to persuade the agent to take an action that increases the principal's payoff or the payoff of a passive third party. We focus on weak punishment that is not sufficient to change the cost of compliance. We do so for two reasons. First, it allows us to focus on the expressive function of punishment through norms, rather than by changing equilibrium behavior. Second, in many real-world situations, punishment is weak due to the high costs of monitoring. To investigate social norms, we follow Bicchieri and Xiao (2009) and Krupka and Weber (2013) and elicit personal norms, injunctive norms and descriptive norms about the agent's behavior in the game.

Consistent with our hypothesis, we find that self-serving punishment sends a weaker

[1] For a review of the experimental literature on punishment, see Xiao (2018).

normative message about the appropriateness of compliance relative to other-regarding punishment. In fact, self-serving punishment actually reduces the social stigma of making the self-interested choice (compared to a scenario in which no punishment is used). In line with these effects on norms, self-serving punishment increases the prevalence of crowding-out, whereby agents who would behave prosocially in the absence of punishment, choose the self-interested action when the principal imposes punishment. This backfiring of punishment is significantly less likely in response to other-regarding punishment.

Our findings have implications for how policymakers, enforcement agencies and institutions should design punishment mechanisms in order to avoid these detrimental crowding-out effects. Specifically, punishment sends a stronger normative signal when agents perceive it to be benefiting others, rather than simply the institution itself. Moreover, we contribute to the existing literature on how punishment affects prosocial behavior and the interplay between punishment mechanisms and social norms, which are increasingly recognized as an important driver of behavior (e.g., Bicchieri and Xiao, 2009; Krupka and Weber, 2013; Gächter et al., 2013; Kimbrough and Vostroknutov, 2016). Our findings shed light on a number of puzzling results from previous studies. For example, punishment has been shown to backfire in the trust game (e.g., Fehr and Rockenbach, 2003), but is often successful at raising contributions in public goods games (e.g., Fehr and Gächter, 2000). Although there are a number of differences between the two games, one key difference is that in trust games punishment only benefits the punisher, while in public goods games punishment can potentially benefit multiple members of the group. Thus, punishment can be perceived as "self-interested" in trust games and as more "other-regarding" in public goods games, which, as our paper shows, has profound implications for the normative message transmitted by punishment.

Recent work has recognized the importance of the norm-transmitting role of punishment and emphasized the benefits of combining punishment with the provision of normative information (e.g., Kölle et al., 2020; Bicchieri et al., 2021).[2] Less is known, however, about which features of punishment can affect the transmission of social norms, and how best to design punishment mechanisms to send a strong normative message. Bowles and Polania-Reyes (2012) emphasise the role of the contextual and institutional details of punishment mechanisms for their effectiveness. For example, punishment can be more effective when it is endogenously chosen by the group (Tyran and Feld, 2006), or implemented in public (Xiao and Houser, 2011). In a related study, Xiao (2013) shows that when punishment results in profits for the punisher, it is less effective in signaling to a third-party whether the

[2] Danilov and Sliwka (2017) study the ability of positive incentives to signal norms and show that the choice of a fixed wage (over a performance-based wage) increases overall effort by changing agents' empirical expectations. See also Van der Weele (2012) on this point.

punishee has lied or told the truth. Our paper differs from Xiao (2013) in that we design and study a context in which the punishee's choice is transparent, but the norm regulating his/her behavior is ambiguous. The design allows us to provide direct evidence on how the punishment motive affects beliefs about norms and the social stigma associated with certain actions. We show that whether the punishment is implemented out of a concern for the punisher or for others can affect the strength of the normative message conveyed and consequently decision-making.

2 Experimental design

We design a simple sequential principal-agent game with three players (Players A, B, and C). Player B ("the agent") chooses between a Communal Project (henceforth, CP) and an Exclusive Project (henceforth, EP). The CP provides the same payoff (£8) to each player. The EP offers a larger benefit to two of the three players (Player B and another player, A or C, depending on treatment – see below) and offers £12 to each of the two "included" players while the "excluded" player receives a lower payoff (£6). Before Player B makes a choice, Player A ("the principal") decides whether to impose a fixed fee to reduce the payoffs of each of the two players who are included in the EP by £2.[3] In our context, the principal imposes a general rule about punishment before knowing the agent's choice, similar to laws which explicitly state the conditions under which fines and other punitive measures will be imposed. The agent is therefore clear on the consequences of their choice when making their decision.

Our two treatments vary whether the player who is excluded from the EP is Player A or Player C ("the third party"). In the *Self* treatment, Player A is the excluded player and receives a higher payoff under CP than EP (see Figure 1). Thus, by imposing the fee, the principal can punish the agent if the agent takes an action (i.e. choosing the EP) that harms the principal. In this sense, punishment is self-serving. In the *Other* treatment, Player C is the excluded player (see Figure 2). By imposing the fee, not only does the principal punish the agent for choosing EP, but also reduces his/her own payoff. In this case, punishment cannot be self-serving and can only benefit the third party.

Note that an important feature of our design is that the two treatments are identical in all aspects (including the agent's incentives), except that in *Self*, punishment can be used to benefit the punisher (Player A), while in *Other* it can only benefit a passive third party (Player C). Moreover, punishment is weak in that the payoffs alone are not sufficient to

[3] The total payoff is higher in the EP than CP both with and without punishment. We chose these payoffs to reflect situations in the real world in which tradeoffs exist between equality and efficiency.

incentivize Player B to change their behavior (Player B always earns more under EP than CP, regardless of whether Player A uses punishment). We elaborate in the next section on the role of punishment in changing the agent's behavior by signaling the underlying norm of conduct. Thus, our treatments shed light on how self-serving motives underlying punishment may influence the perception of social norms and hence behavior.

Figure 1: *Self* treatment

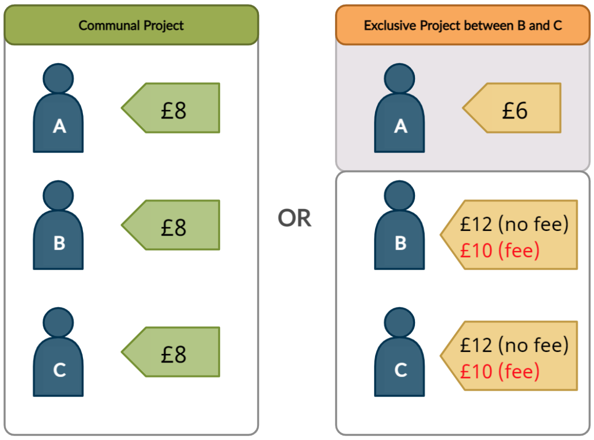
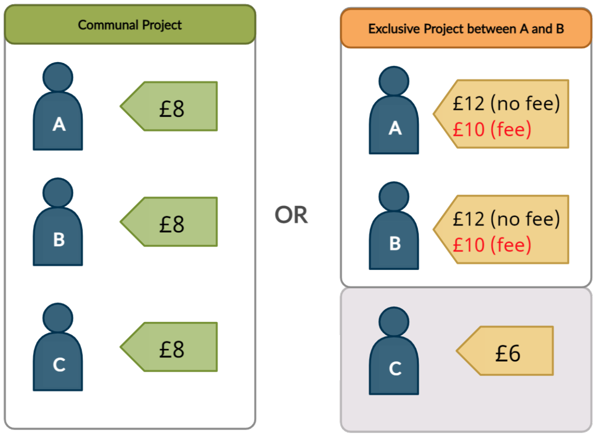


Figure 2: *Other* treatment



In each treatment, Player A was asked to make a decision about whether to use punishment or not. We elicited Player B's decisions using a strategy elicitation method, i.e., we asked Player B to make one choice in case A imposed a fee, and one choice in case A did not impose a fee.[4] Our analysis, which we pre-registered together with the experimental design on AsPredicted.org (pre-registration #64211), will focus on how the agent's strategies change based on the principal's punishment decision. In particular, depending on the agent's choices, we classify them as one of four possible types: (i) "Unconditional CP" if they choose the CP regardless of whether A uses punishment; (ii) "Unconditional EP" if they choose the EP regardless of punishment; (iii) "Crowded-in" if they choose the CP when A uses punishment and the EP when A does not; and (iv) "Crowded-out" if they perversely choose the EP when A uses punishment and the CP when A does not. Our key question is whether the motive behind punishment affects the distribution of B's types across the two treatments, and in particular, the share of subjects who are Crowded-out types.

The other key focus of the paper is on how punishment affects social norms across the two treatments. We elicited social norms from subjects assigned to the role of Player C, *before* we actually revealed their role to them, so that their normative beliefs would not be

[4] We randomised the order in which we elicited these two choices to control for possible order effects.

biased by any player-specific considerations.[5] These subjects were asked to answer a few questions about the behavior of previous participants in the task before being informed of their role.[6] After answering the questions in the first part, subjects moved to the second part, where there were told they would participate in the game (either *Self* or *Other*, depending on the treatment) they had just evaluated, in the role of Player C.

The norm-elicitation questions are based on the Bicchieri and Xiao (2009) procedure to elicit social norms.[7] We first asked participants for their first-order beliefs about the appropriateness of choosing the EP and the CP, with and without punishment (four questions in total). Subjects indicated their judgment using a 5-point scale ranging from "Very appropriate" to "Very inappropriate", and were told that by "appropriate" we meant behavior that they "personally believe is the correct or ethical thing to do". These first-order beliefs were not incentivized and can be interpreted as how participants personally felt about the appropriateness of each choice, or their *personal norms*, which may or may not align with the perceived views of the majority.[8]

Second, we elicited subjects' second-order beliefs by asking them to guess the most common first-order beliefs of participants in a previous session (the pilot experiment mentioned in footnote 6). We elicited a second-order belief in correspondence to each of the four first-order beliefs discussed above (appropriateness of choosing EP when A punishes; appropriateness of choosing CP when A punishes; appropriateness of choosing EP when A does not punish; appropriateness of choosing CP when A does not punish). Again, subjects indicated their responses on a 5-point scale ranging from "Very appropriate" to "Very inappropriate". We incentivized these responses by paying participants an additional £1 if their guess was correct for one of the four questions, randomly chosen. Since these guesses measure subjects' beliefs of what others consider appropriate or inappropriate, they express subjects' perception of the *injunctive norm* that surrounds B's behavior in the game.

[5] There is mixed evidence regarding whether player-specific considerations affect elicited norms. Erkut et al. (2015) find little evidence that this is the case in a dictator game, but Heinicke et al. (2022) find the opposite result in a series of mini-dictator games with moral wiggle room. We did not elicit norms from Players A and B before informing them of their role because we were worried that merely asking them to think about social norms may have altered their subsequent game behavior. This is known as the "focusing effect" of norms whereby focusing a decision-maker's attention on norms can activate norm compliance (e.g., Krupka and Weber, 2009; d'Adda et al., 2016)

[6] These previous participants were subjects recruited to take part in a pilot (N=120) that we used to conduct a power analysis to calibrate the study's sample size. The pilot was identical to the main experiment, except that Player C's were only asked unincentivized questions. We used the data from the pilot to incentivize Player C's answers in the main experiment.

[7] See also Krupka and Weber (2013) for a related norm-elicitation procedure and Görges and Nosenzo (2020) for a review of the experimental literature on the elicitation of norms.

[8] Bašić and Verrina (2021) show that personal norms can differ from social norms (second-order beliefs) and are predictive of behavior.

Finally, we elicited subjects' empirical beliefs by asking them to guess the percentage of Player B's in a previous session (the pilot experiment) who actually chose the EP (by construction, the remainder would have chosen the CP), under punishment and under no punishment (two questions in total). These questions measure subjects' perception of the *descriptive norm* of behavior in the game. We incentivized empirical beliefs using the Karni (2009) mechanism, a variation of the Becker-DeGroot-Marschak procedure (Becker et al., 1964).[9] Descriptive norms can differ from injunctive norms and can be particularly useful in explaining behavior when an injunctive norm is not followed in practice (e.g., Bicchieri and Xiao, 2009).

The experiment was programmed in oTree (Chen et al., 2016) and was conducted on Prolific in April 2021 (see Appendix A for screenshots of players' decision screens). We randomly matched three participants to form a group and randomly assigned each participant to one of the three roles in the game (A, B or C). Subjects were randomly assigned to a treatment (either *Self* or *Other*). We report data from N=883 participants with N=425 in *Self* and N=458 in *Other*.[10] The sample size was determined based on a power analysis conducted after we ran a small pilot with 60 subjects per treatment. In the pilot we observed a treatment effect on the distribution of types of size 0.33 (Cohen's d). We chose a sample of 150 subjects per role per treatment to be able to detect at least 75% of the effect size observed in the pilot (i.e., Cohen's d = 0.24), with 95% power and alpha = 0.05. To improve data quality and homogeneity, we restricted participation to individuals residing in the United Kingdom, with an approval rate higher than 80% on Prolific. Participants received a completion fee of £1.50 and we selected 1 in every 20 participants to receive their earnings from the game as a bonus payment, as well as payments based on their second-order normative beliefs and empirical beliefs (if applicable). Decisions were anonymous and participants earned an average of £2.60 for a median completion time of 7.5 minutes.

[9] We chose the Karni mechanism due to its invariance to heterogeneous risk preferences. See Schwardmann and van der Weele (2019) for a similar elicitation question, presented as a multiple price list. Following Danz et al. (2020) who find that belief accuracy is higher with less information about the payment mechanism, we informed participants that their chances of receiving an additional £1 were highest when they made their "best guess" and gave the option to separately see more details about the payment mechanism if they wished (58% chose to do so).

[10] As specified in our pre-registration, we exclude from our analysis 49 participants who did not correctly answer all of the control questions (after two attempts). Our main results remain unchanged with the inclusion of these participants.

3 Theoretical framework and hypotheses

In this section, we present a theoretical framework to derive our hypotheses. Our main research question is whether the motive behind punishment affects (1) the normative message conveyed, and (2) the agent's actual behavior in the game.

If the agent only cares about maximizing material payoffs, in both versions of the game they have a dominant strategy to choose the EP, regardless of the punishment decision of the principal. Anticipating this, the principal chooses not to punish in *Other*, and is indifferent between punishing or not in *Self*.

Literature in behavioral economics has documented that agents care about more than material payoffs. We adopt a norm-based utility function framework in which utility depends on material payoffs and norm compliance: agents experience a disutility when they violate a social norm, due to the social disapproval or stigma they receive for breaking the norm (e.g., Bicchieri, 2005; Krupka and Weber, 2013).[11] We further assume that, in the context of the game studied here, the norm prescribes that Player B chooses the CP.[12] When a player chooses the EP, they experience a disutility equal to the (positive) difference in appropriateness between choosing the CP and choosing the EP. The larger this difference, the stronger the relative stigma for choosing the EP over the CP. Crucially, below we will assume that the strength of this stigma depends on whether choosing the EP incurs punishment.

In the game, the agent chooses one action under no punishment ($a_{NoPun} \in \{CP, EP\}$), and one action under punishment ($a_{Pun} \in \{CP, EP\}$). Without punishment, the agent receives $\pi(CP) = 8$ and $\pi(EP) = 12$. The principal decides whether to impose a fee $f \in \{0, 2\}$, which is implemented only if the agent chooses $a_{Pun} = EP$.

Let $k > 0$ represent the agent's sensitivity towards norms and $S \geq 0$ the relative stigma for choosing the EP instead of the CP. The agent's *net* utility for choosing the EP instead of the CP is therefore given by: $4 - f - k \cdot S$. We now analyze the agent's best-response to the principal's punishment decision, as a function of k and S.

Case 1: If the principal does not punish ($f = 0$), the agent's best-response is:

$$\begin{cases} a^*_{NoPun} = CP, & \text{if } S_{NoPun} \geq 4/k \\ a^*_{NoPun} = EP, & \text{otherwise} \end{cases} \tag{1}$$

[11] One interpretation of this disutility is that the stigma of norm violation represents the costs of deviating from norms in repeated interactions (Binmore, 2005).

[12] Our norms data indeed confirms this since in all elicitations the appropriateness of choosing the CP is greater than the appropriateness of choosing the EP (see Appendix B).

For a given norm sensitivity parameter k, the greater the relative stigma of choosing the EP instead of the CP, the more likely it is that the agent chooses the CP. Similarly, the higher is k, the more likely it is that the agent chooses the CP, ceteris paribus.

Case 2: If the principal does punish ($f = 2$), the agent's best-response is:

$$\begin{cases} a^*_{NoPun} = CP, & \text{if } S_{Pun} \geq 2/k \\ a^*_{NoPun} = EP, & \text{otherwise} \end{cases} \tag{2}$$

As before, the agent's choice depends on the size of the relative stigma against the EP and the agent's norm sensitivity parameter. However, because the principal has imposed a fee, which makes the EP less attractive in monetary terms for the agent, the threshold values of S_{Pun} and k are lower than under the case of no punishment.

Taken together, these conditions define the threshold values of S and k that determine the agent's best-response strategy. There are four cases:

$$\{a^*_{NoPun}, a^*_{Pun}\} = \begin{cases} \{CP, CP\}, & \text{if } S_{NoPun} \geq 4/k, S_{Pun} \geq 2/k \\ \{EP, EP\}, & \text{if } S_{NoPun} < 4/k, S_{Pun} < 2/k \\ \{EP, CP\}, & \text{if } S_{NoPun} < 4/k, S_{Pun} \geq 2/k \\ \{CP, EP\}, & \text{if } S_{NoPun} \geq 4/k, S_{Pun} < 2/k \end{cases} \tag{3}$$

These four cases correspond to the four agent types that we defined in Section 2 (Unconditional CP; Unconditional EP; Crowded-in; Crowded-out). The framework clarifies that the relative frequency of each type depends on the distribution of the norm sensitivity parameter and the relative stigma against the EP, which we assume is affected by punishment.

Therefore, our first hypothesis concerns the effect that punishment has on the relative stigma against the EP. We conjecture that punishment that is devoid of self-serving motives sends a stronger normative message regarding what is considered appropriate behavior and therefore triggers a relatively stronger change in the stigma against the EP relative to the case without punishment. In particular, let ΔS^{Self} be the difference between S_{Pun} and S_{NoPun} in the *Self* treatment, and ΔS^{Other} be the difference in the *Other* treatment. We conjecture that ΔS^{Other} is likely to be positive since choosing the EP is likely to trigger strong stigma especially when a principal is willing to reduce his/her own payoffs to impose a fee when the agent's choice harms a third party. On the other hand, the effect may be smaller in the *Self* treatment, where the normative message of punishment may be "diluted" by the fact that the principal has a direct interest at stake in the choice of the agent. In fact, if punishment is perceived as self-servingly coercive (after all, choosing the EP maximizes joint profits and makes the agent and the third party better off), ΔS^{Self} may even be negative, i.e. punishment may reduce the stigma against the EP if choosing the EP is seen as a legitimate form of retaliation against self-serving punishment. We summarize these considerations in the following pre-registered hypothesis:

Hypothesis 1: Other-regarding punishment increases the stigma against choosing the EP more than self-serving punishment.

$$\Delta S^{Self} < \Delta S^{Other} \tag{4}$$

If our first hypothesis is confirmed, this can have direct implications for the distribution of agents' types we should observe across the two treatments. In particular, if $\Delta S > 0$, there cannot be Crowded-out agents, because this type only emerges when the stigma against the EP, for any given k, is relatively larger under no punishment than under punishment (i.e., when $\Delta S < 0$; see (3) above and also Figure C.1 in Appendix C). Thus, if Hypothesis 1 is confirmed and $\Delta S^{Other} > 0 > \Delta S^{Self}$, then we expect self-serving punishment to be more likely to backfire than other-regarding punishment. We summarize these considerations in our second pre-registered hypothesis:

Hypothesis 2: Punishment is more likely to backfire (i.e., induce more Crowded-out types) when it is motivated by self-interest compared to when it is motivated by other-regarding concerns.

4 Results

The focus of this section is to study how punishment affects the normative message of punishment and its effectiveness. Overall, principals use punishment more often in *Self* (48.6%) than in *Other* (24.5%) and this difference is significant according to a χ^2 test ($p < 0.01$) . In Section 4.1 we investigate how punishment affects the stigma for choosing the EP (Hypothesis 1). In Section 4.2 we examine agents' choices and the effectiveness of punishment (Hypothesis 2).

4.1 The normative message conveyed by punishment

We study Hypothesis 1 by inspecting how punishment affects the relative stigma against the EP. Note that we have collected social norms data using three different norm-elicitation questions, pertaining to first-order beliefs of appropriateness (personal norm), second-order beliefs of social appropriateness (injunctive norm) and first-order beliefs of the frequency of agents' choices (descriptive norm). We can thus construct three distinct measures of stigma, based on personal norms, injunctive norms and descriptive norms. Table 1 reports data from these norm-elicitations. The table reports both the average absolute levels of S_{Pun} and S_{NoPun} across our treatments, as well as the resulting values of ΔS.[13]

Punishment in *Self* reduces the relative stigma against the EP across all three norm measures. The drop in stigma is statistically significant for personal norms ($p < 0.01$; two-tailed Wilcoxon signed-rank test) and injunctive norms ($p < 0.01$).[14] The drop is instead insignificant for descriptive norms ($p = 0.56$). In contrast, punishment does not significantly change personal norms in *Other* ($p = 0.87$), but does increase relative stigma for the injunctive norm ($p = 0.04$), as well as for the descriptive norm ($p < 0.01$). Thus, in line with our conjectures, $\Delta S^{Self} \leq 0$, while $\Delta S^{Other} \geq 0$.

Table 1: Stigma of choosing the EP

	Personal norm			Injunctive norm			Descriptive norm		
	S_{NoPun}	S_{Pun}	ΔS	S_{NoPun}	S_{Pun}	ΔS	S_{NoPun}	S_{Pun}	ΔS
Self	1.75 (1.77)	0.73 (1.75)	**-1.02** **(2.06)**	1.76 (1.98)	0.62 (2.08)	**-1.14** **(2.87)**	42.45 (30.50)	40.38 (28.70)	**-2.07** **(47.18)**
Other	1.44 (1.76)	1.38 (1.59)	**-0.06** **(1.68)**	1.11 (2.00)	1.41 (1.83)	**0.30** **(2.12)**	29.66 (23.97)	51.26 (25.32)	**21.60** **(36.90)**

Notes: For personal and injunctive norms, in line with our theoretical framework, S is calculated as: (appropriateness of choosing CP) - (appropriateness of choosing EP). For descriptive norms, our measurement of S is simply the expected percentage of CP choices (note that this is a departure from our definition of S in the theoretical framework; adapting the framework to the empirical measure is however straightforward). ΔS is calculated as: S_{Pun} - S_{NoPun}. A positive value means punishment increases the stigma of choosing EP, while a negative value means punishment reduces the stigma. Standard deviations in parentheses.

We test Hypothesis 1 by comparing ΔS^{Self} and ΔS^{Other} for each of our norm measures. We find that other-regarding punishment increases the stigma against the EP more than self-serving punishment, both when we look at personal norms (-1.02 vs. -0.06, $p < 0.01$;

[13] See Appendix B for the appropriateness ratings for personal and injunctive norms.

[14] Unless otherwise stated, we use two-tailed Wilcoxon signed-rank tests to compare changes in stigma due to punishment.

two-tailed Mann-Whitney test) and injunctive norms (-1.14 vs. 0.30, $p < 0.01$).[15] Moreover, subjects expect a larger increase in CP choices in response to other-regarding punishment as compared to self-serving punishment (-2.07 vs. 21.60, $p < 0.01$). These findings are corroborated by the regression analysis (which also controls for demographic variables), presented in Table 2. Thus, this analysis confirms our first hypothesis, as we summarize in the following result:

Table 2: How punishment changes the stigma against the EP (ΔS)

	Personal norm		Injunctive norm		Descriptive norm	
	(1)	(2)	(3)	(4)	(5)	(6)
Self	-0.96^{***}	-0.93^{***}	-1.45^{***}	-1.52^{***}	-23.67^{***}	-25.15^{***}
	(0.22)	(0.23)	(0.29)	(0.32)	(4.94)	(5.26)
Constant	-0.06	0.05	0.30	1.73	21.60^{***}	41.27^{**}
	(0.15)	(0.85)	(0.20)	(1.17)	(3.43)	(19.38)
Controls	No	Yes	No	Yes	No	Yes
R^2	0.06	0.16	0.08	0.12	0.07	0.15
Adj. R^2	0.06	0.08	0.07	0.04	0.07	0.06
Num. obs.	292	291	292	291	292	291

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: OLS regression with standard errors in parentheses. The dependent variable is ΔS, computed using first-order beliefs of personal norms (Columns 1 and 2), second-order beliefs of injunctive norms (Columns 3 and 4) and first order beliefs of descriptive norms (Columns 5 and 6). The baseline treatment is *Other*. The control variables are the order in which agents' choices were elicited, gender, age, education, religiosity, income and political orientation.

Result 1: Consistent with Hypothesis 1, other-regarding punishment increases the relative stigma against the EP more than self-serving punishment.

4.2 The effectiveness of punishment

The previous section showed that there is a fundamental difference between self-serving and other-regarding punishment. The former reduces the stigma against choosing selfish behavior, while the latter strengthens it. We now assess whether these differences in the normative message transmitted by punishment translate into actual behavioral differences.

We first examine agents' choices in the two treatments, based on whether the principal chose to punish or not. In *Self*, 47.2% of agents choose the CP in the absence of punishment,

[15] Unless otherwise stated we use two-tailed Mann-Whitney tests to compare the change in stigma across *Self* and *Other* for each norm measure.

and 47.9% choose the CP with punishment (McNemar's test, $p = 1.00$). In *Other*, 38.5% choose the CP under no punishment, while 55.4% do so under punishment (McNemar's test, $p < 0.01$). Table 3 similarly shows that when punishment is imposed in *Other*, it is 2.16 times ($p < 0.01$, column 4) more likely that agents will choose the CP, while in *Self* choices are not significantly different when punishment is used ($p = 0.90$, column 2).[16] Our findings suggest that punishment is effective at changing behavior, but only when it is motivated by a concern for others.

Table 3: Likelihood of choosing the CP

	Self		*Other*	
	(1)	(2)	(3)	(4)
Pun	1.029	1.031	1.984***	2.155***
	(0.177)	(0.194)	(0.160)	(0.178)
Constant	0.893	0.591	0.626***	1.037
	(0.168)	(1.273)	(0.169)	(1.090)
Controls	No	Yes	No	Yes
AIC	397.00	414.31	404.73	418.76
BIC	404.30	501.88	412.11	514.71
Log Likelihood	−196.50	−183.15	−200.36	−183.38
Deviance	393.00	366.31	400.73	366.76
Num. obs.	284	284	296	296

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: Odds ratio logistic regression with standard errors clustered at the individual level in parentheses. The dependent variable is the agent's choice (= 1 if they chose CP). The control variables are order in which the agent's choice was elicited, gender, age, education, religiosity, income and political orientation.

To test Hypothesis 2, we compare the distribution of agents' types between the two treatments. Across *Self* and *Other*, we find a similar share of Unconditional CP (33.8% vs. 34.5%) and Unconditional EP (38.7% vs. 40.5%) types. It is not surprising that these two types represent a majority of agents in our sample given that we examine a weak form of punishment.[17] We observe a smaller proportion of Crowded-in types (for whom punishment induced a switch from the EP under no punishment to the CP under punishment) in *Self* than in *Other* (14.1% vs. 20.9%). Conversely, we find a larger proportion of Crowded-out types (for whom punishment backfired) under self-serving punishment, compared to other-regarding punishment (13.4% vs. 4.1%). According to a χ^2 test, the distribution of types

[16] We find no evidence of an order effect, see Appendix D.

[17] Another possibility is that the use of a strategy elicitation means we are more likely to observe consistency in agents' choices and might underestimate the number of Crowded-in and Crowded-out types. Our goal is not to draw conclusions about the levels of compliance or non-compliance, but rather to compare the relative effectiveness of punishment, given different underlying motivations.

across *Self* and *Other* is significantly different ($p = 0.03$).

This result is also supported by the multinomial logistic regression analysis in Table 4, which compares the likelihood of observing each agent type against each of the other agent types under self-serving punishment, relative to other-regarding punishment. Columns 1-3 compare the likelihood of observing the Unconditional CP type against Unconditional EP, Crowded-in and Crowded-out types. In columns 4-5, we present the likelihood of the Unconditional EP type against Crowded-in and Crowded-out types. Column 6 compares the likelihood of observing the Crowded-in type relative to the Crowded-out type. Relative to *Other*, agents in *Self* are 2.83 times more likely to be a Crowded-out type than an Unconditonal CP type ($p = 0.06$, column 3) and 2.96 times more likely to be a Crowded-out type than an Unconditional EP type ($p = 0.01$, column 5). In *Self*, we are also 4.28 times more likely to observe a Crowded-out type than a Crowded-in type, relative to *Other* ($p = 0.04$, column 6). The relative shares of Unconditional CP, Unconditional EP and Crowded-in types against one another are instead unchanged across the two treatments. These results confirm Hypothesis 2 and show that punishment is more likely to backfire when it is motivated by self-interest than by other-regarding motives, as we summarize in the following result.

Table 4: Likelihood of observing agents' types

	Uncond CP			Uncond EP		Crowd-in
	Uncond_EP	Crowd-in	Crowd-out	Crowd-in	Crowd-out	Crowd-out
	(1)	(2)	(3)	(4)	(5)	(6)
Self	0.956	0.662	2.830*	0.692	2.961**	4.277**
	(0.294)	(0.377)	(0.548)	(0.366)	(0.541)	(0.586)
Constant	0.910	0.406	0.128	0.447	0.140	0.315
	(1.019)	(1.632)	(2.102)	(1.610)	(2.059)	(2.431)

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: Odds ratio multinomial logistic regression with standard errors in parentheses (N=290, AIC: 800.378). The dependent variable is agent's type based on their choices. The baseline treatment is *Other*. The control variables are the order in which agents' choices were elicited, gender, age, education, religiosity, income and political orientation. Created using the Stargazer package (Hlavac, 2013) in R.

Result 2: Consistent with Hypothesis 2, self-serving punishment is more likely to backfire and crowd out norm compliance compared to other-regarding punishment. Specifically, it increases the share of agents who react perversely to punishment (Crowded-out) compared to all other types of agents.

5 Conclusion

Punishment can be effective at encouraging prosocial behavior. However, the specific factors which lead to punishment crowding out or crowding in prosocial choices remain an open question. We investigate whether the perceived motive behind a punishment decision changes the normative message that is conveyed. We conjecture that punishment that is motivated by self-serving concerns is less effective at reigning in self-interest than punishment that is perceived to be motivated by other-regarding concerns.

Our key takeaways can be summarized as follows. First, by eliciting perceptions of norms (personal, injunctive and descriptive), we find that other-regarding punishment increases the social stigma against self-interested choices, while self-serving punishment can have a detrimental effect by reducing this stigma. Second, consistent with these changes in social stigma and in line with a simple theoretical framework, when punishment is self-serving in nature, agents tend to respond in a perverse manner – by acting more prosocially when punishment is not used than when it is used. Punishment therefore backfires as agents respond to self-serving punishment by also pursuing their own self-interest. Conversely, punishment motivated by other-regarding concerns is effective at encouraging prosocial behavior.

Our results show that, in order for punishment mechanisms to be effective at constraining self-interest, punishment needs to communicate a strong normative message, and that the strength of this message crucially depends on the perceived motives behind punishment choices. Our findings have useful applications for the design of punishment mechanisms, and especially for mechanisms that are monetary in nature, such as fines and taxes. Our results caution that such mechanisms should be designed in a way that clearly communicates the benefits to the wider community (or a specific third party) and minimizes the chances that punishment is interpreted as a profit-making device, or used purely to benefit the enforcement agency.

This paper also sheds light on why punishment is generally effective at constraining self interest in public goods games when it can benefit multiple individuals, but tends to backfire in trust games when it is used only to benefit the punisher. A promising avenue for future work is to examine other differences between the two punishment contexts which could affect the normative message that is conveyed by punishment, such as the number of potential benefactors of punishment and the nature of the punishment institution.

References

Andreoni, J., Harbaugh, W., and Vesterlund, L. (2003). The carrot or the stick: Rewards, punishments, and cooperation. *American Economic Review*, 93(3):893–902.

Bašić, Z. and Verrina, E. (2021). Personal norms—and not only social norms—shape economic behavior. *MPI Collective Goods Discussion Paper*, (2020/25).

Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral science*, 9(3):226–232.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

Bénabou, R. and Tirole, J. (2011). Laws and norms.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms.* Cambridge University Press.

Bicchieri, C., Dimant, E., and Xiao, E. (2021). Deviant or wrong? The effects of norm information on the efficacy of punishment. *Journal of Economic Behavior and Organization*, 188:209–235.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208.

Binmore, K. (2005). *Natural justice.* Oxford university press.

Bowles, S. and Polania-Reyes, S. (2012). Economic incentives and social preferences: substitutes or complements? *Journal of Economic Literature*, 50(2):368–425.

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

d'Adda, G., Drouvelis, M., and Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, 62:1–7.

Danilov, A. and Sliwka, D. (2017). Can contracts signal social norms? Experimental evidence. *Management Science*, 63(2):459–476.

Erkut, H., Nosenzo, D., and Sefton, M. (2015). Identifying social norms using coordination games: Spectators vs. stakeholders. *Economics Letters*, 130:28–31.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Fehr, E. and Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422(6928):137–140.

Gächter, S., Nosenzo, D., and Sefton, M. (2013). Peer effects in pro-social behavior: Social norms or social preferences? *Journal of the European Economic Association*, 11(3):548–573.

Galbiati, R., Schlag, K. H., and van der Weele, J. J. (2013). Sanctions that signal: An experiment. *Journal of Economic Behavior and Organization*, 94:34–51.

Gneezy, U. and Rustichini, A. (2000). Pay Enough or Don't Pay at All. *The Quarterly Journal of Economics*, 115(3):791–810.

Görges, L. and Nosenzo, D. (2020). Measuring social norms in economics: Why it is important and how it is done. *Analyse & Kritik*, 42(2):285–311.

Heinicke, F., König-Kersting, C., and Schmidt, R. (2022). Injunctive vs. descriptive social norms and reference group dependence. *Journal of Economic Behavior and Organization*, 195:199–218.

Hlavac, M. (2013). stargazer: Latex code and ascii text for well-formatted regression and summary statistics tables. *URL: http://CRAN. R-project. org/package= stargazer.*

Kahan, D. M. (1998). Social meaning and the economic analysis of crime. *The Journal of Legal Studies*, 27(S2):609–622.

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

Kimbrough, E. O. and Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 14(3):608–638.

Kölle, F., Lane, T., Nosenzo, D., and Starmer, C. (2020). Promoting voter registration: the effects of low-cost interventions on behaviour and norms. *Behavioural Public Policy*, 4(1):26–49.

Krupka, E. and Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, 30(3):307–320.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

McAdams, R. H. (2000). A focal point theory of expressive law. *Virginia Law Review*, pages 1649–1729.

Posner, R. A. (1997). Social norms and the law: An economic approach. *The American Economic Review*, 87(2):365–369.

Schwardmann, P. and van der Weele, J. (2019). Deception and self-deception. *Nature Human Behaviour*, 3(10):1055–1061.

Sunstein, C. R. (1996). On the expressive function of law. *University of Pennsylvania law review*, 144(5):2021–2053.

Tyran, J. R. and Feld, L. P. (2006). Achieving compliance when legal sanctions are non-deterrent. *Scandinavian Journal of Economics*, 108(1):135–156.

Van der Weele, J. (2012). The signaling power of sanctions in social dilemmas. *The Journal of Law, Economics, & Organization*, 28(1):103–126.

Villatoro, D., Andrighetto, G., Brandts, J., Nardin, L. G., Sabater-Mir, J., and Conte, R. (2014). The norm-signaling effects of group punishment: combining agent-based simulation and laboratory experiments. *Social Science Computer Review*, 32(3):334–353.

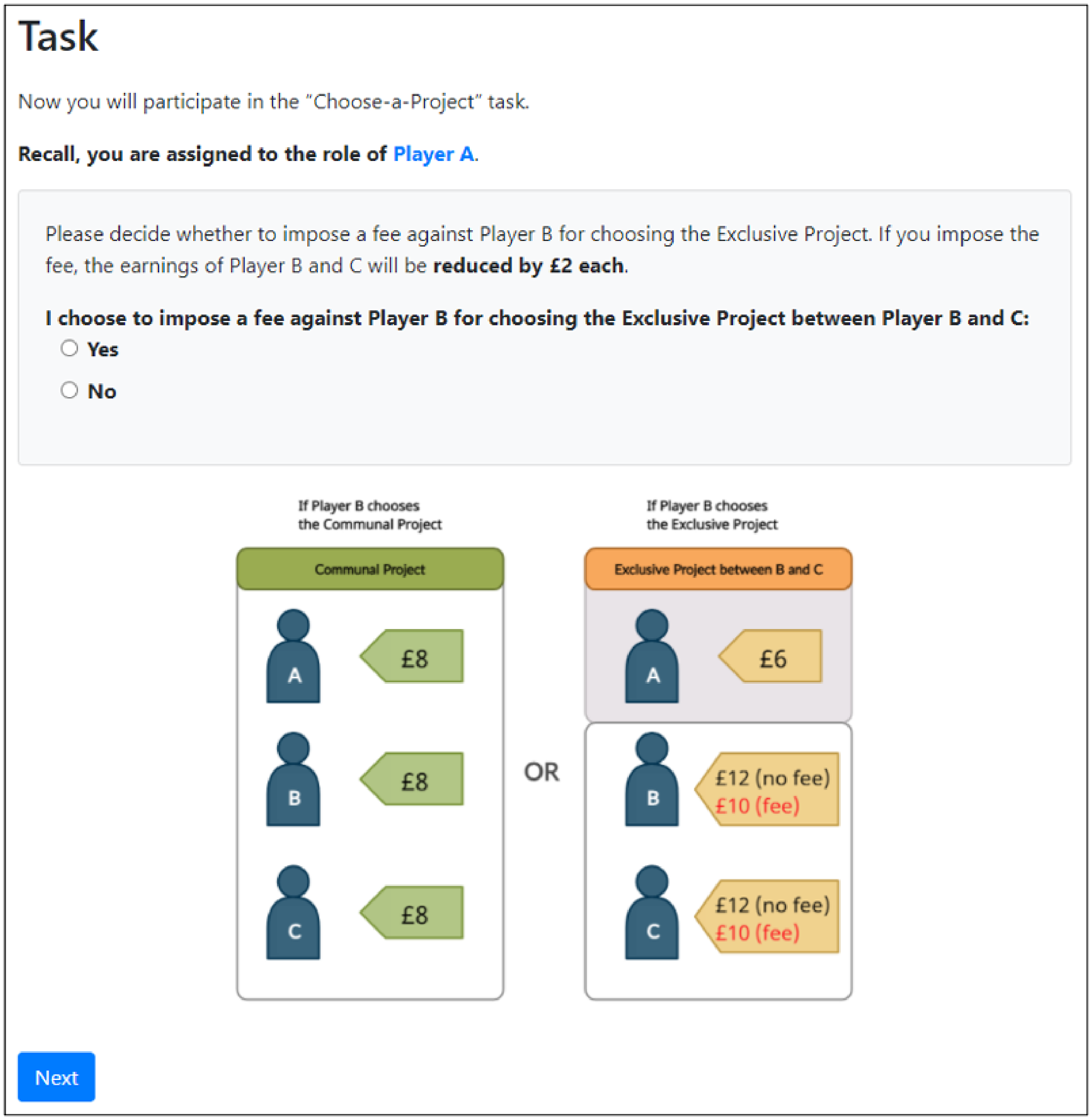
Xiao, E. (2013). Profit-seeking punishment corrupts norm obedience. *Games and Economic Behavior*, 77(1):321–344.

Xiao, E. (2018). Punishment, social norms, and cooperation. In *Research Handbook on Behavioral Law and Economics*. Edward Elgar Publishing.

Xiao, E. and Houser, D. (2011). Punish in public. *Journal of Public Economics*, 95(7-8):1006–1017.

A Instructions

Figure A.1: The principal's choice



Task

Now you will participate in the "Choose-a-Project" task.

Recall, you are assigned to the role of Player A.

Please decide whether to impose a fee against Player B for choosing the Exclusive Project. If you impose the fee, the earnings of Player B and C will be **reduced by £2 each**.

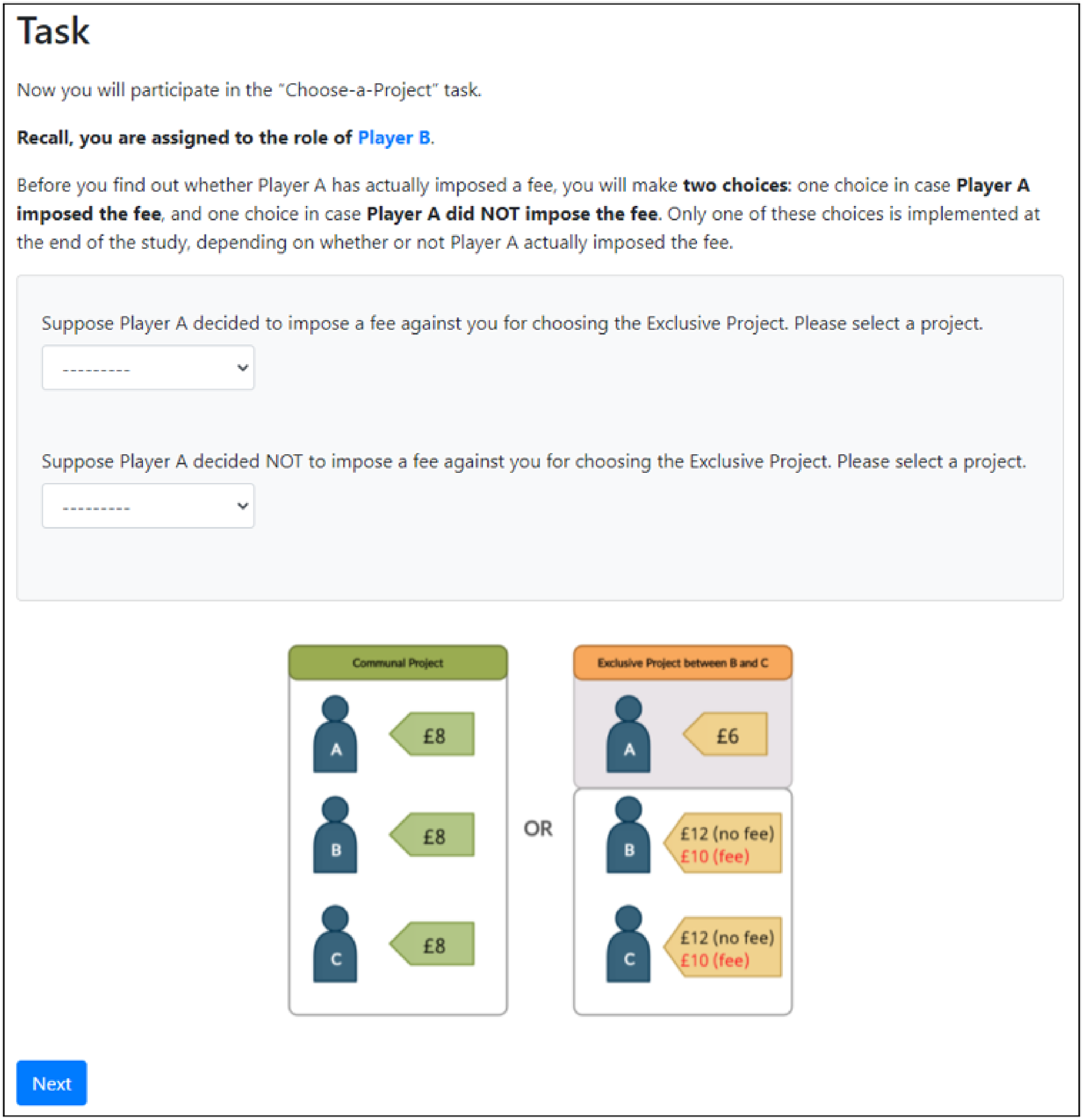
I choose to impose a fee against Player B for choosing the Exclusive Project between Player B and C:

○ **Yes**

○ **No**

If Player B chooses the Communal Project
Communal Project
A
£8
B
£8
C
£8
OR
If Player B chooses the Exclusive Project
Exclusive Project between B and C
A
£6
B
£12 (no fee)
£10 (fee)
C
£12 (no fee)
£10 (fee)
Next

Figure A.2: The agent's choice (Order 1: Pun, NoPun)



Task

Now you will participate in the "Choose-a-Project" task.

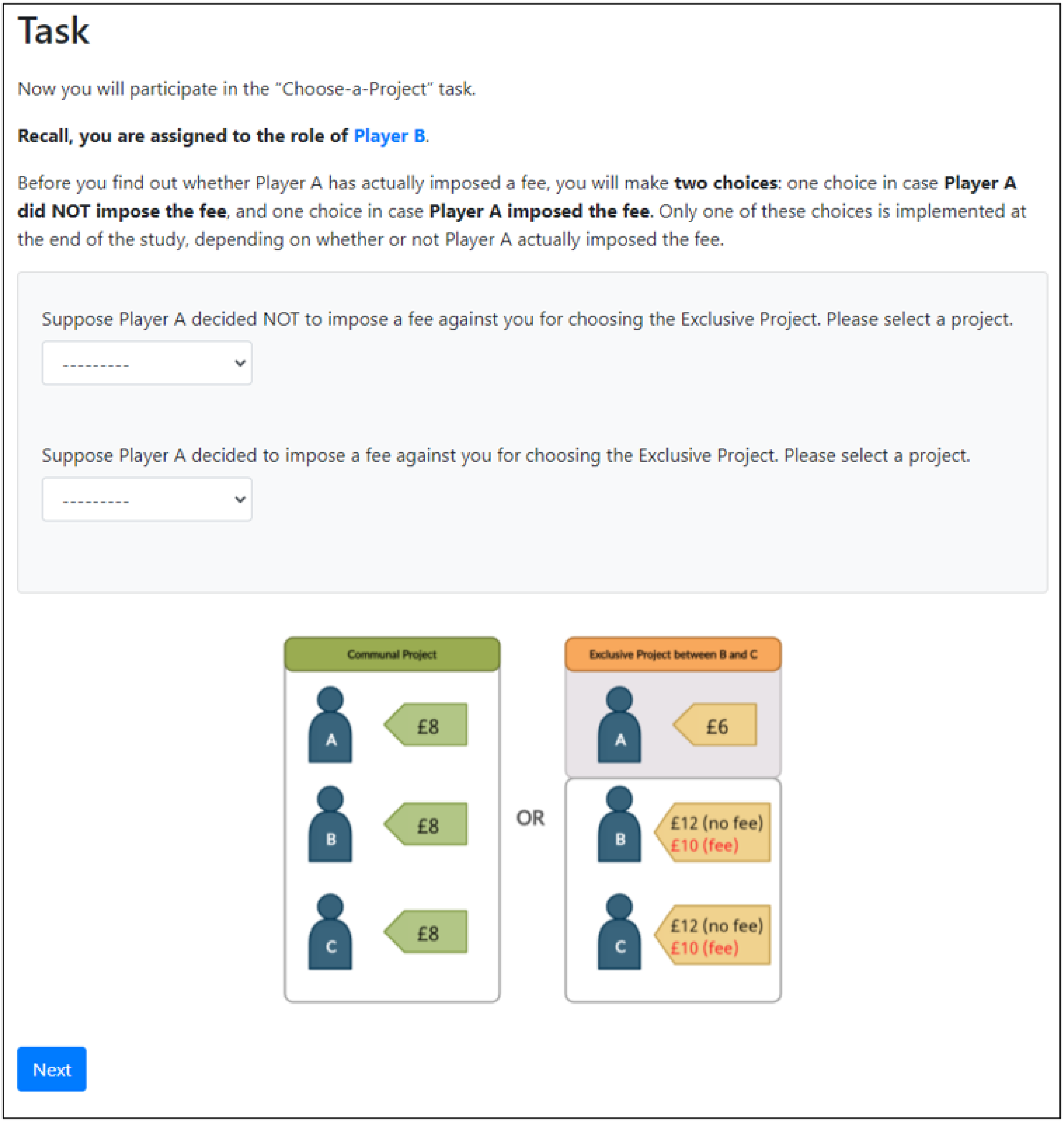
Recall, you are assigned to the role of Player B.

Before you find out whether Player A has actually imposed a fee, you will make **two choices**: one choice in case **Player A imposed the fee**, and one choice in case **Player A did NOT impose the fee**. Only one of these choices is implemented at the end of the study, depending on whether or not Player A actually imposed the fee.

Suppose Player A decided to impose a fee against you for choosing the Exclusive Project. Please select a project.

Suppose Player A decided NOT to impose a fee against you for choosing the Exclusive Project. Please select a project.

Figure A.3: The agent's choice (Order 2: NoPun, Pun)



Task

Now you will participate in the "Choose-a-Project" task.

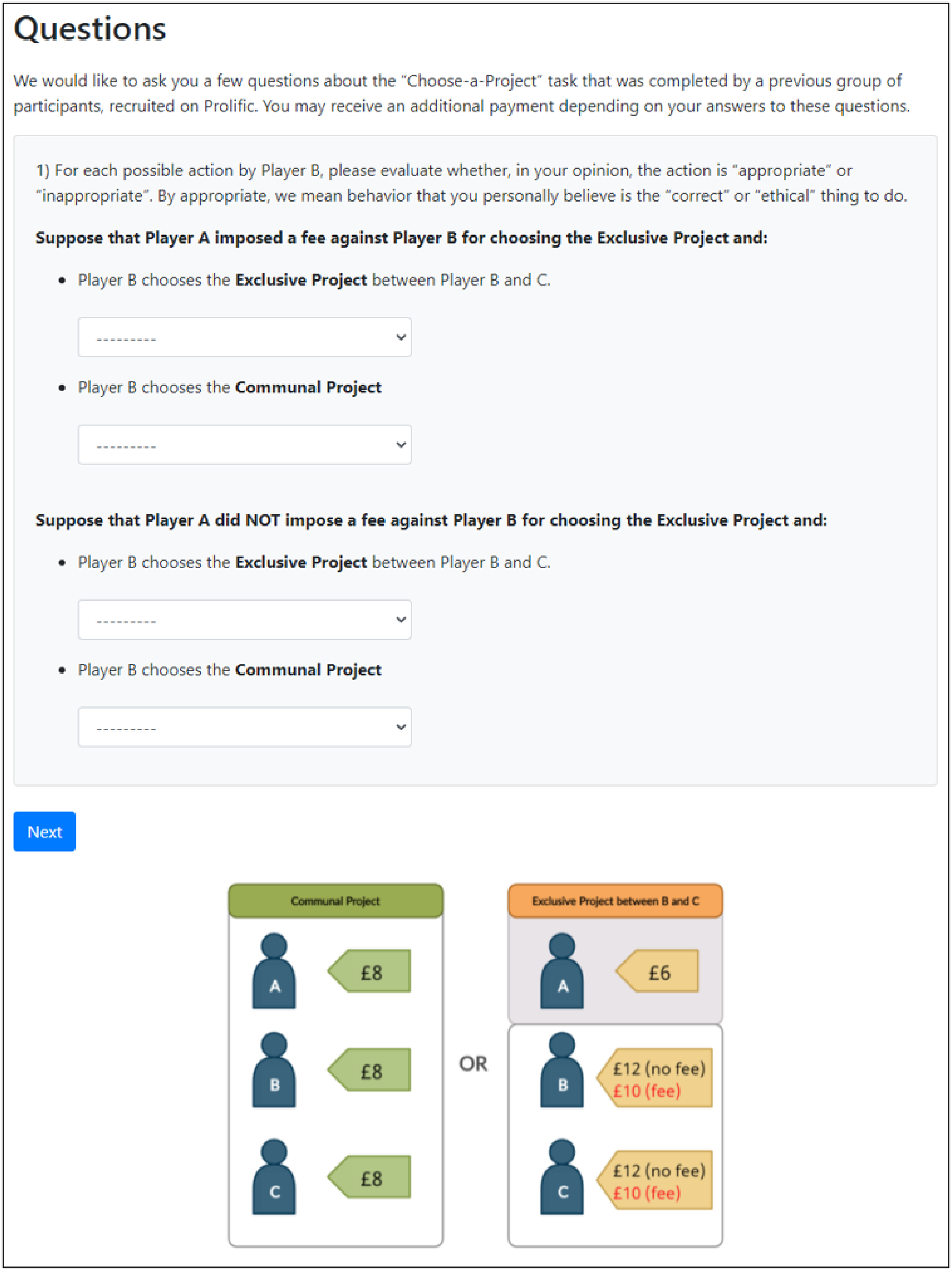
Recall, you are assigned to the role of Player B.

Before you find out whether Player A has actually imposed a fee, you will make **two choices**: one choice in case **Player A did NOT impose the fee**, and one choice in case **Player A imposed the fee.** Only one of these choices is implemented at the end of the study, depending on whether or not Player A actually imposed the fee.

Suppose Player A decided NOT to impose a fee against you for choosing the Exclusive Project. Please select a project.

Suppose Player A decided to impose a fee against you for choosing the Exclusive Project. Please select a project.

Figure A.4: Eliciting third-party personal norms (Order 1)



Questions

We would like to ask you a few questions about the "Choose-a-Project" task that was completed by a previous group of participants, recruited on Prolific. You may receive an additional payment depending on your answers to these questions.

1) For each possible action by Player B, please evaluate whether, in your opinion, the action is "appropriate" or "inappropriate". By appropriate, we mean behavior that you personally believe is the "correct" or "ethical" thing to do.

Suppose that Player A imposed a fee against Player B for choosing the Exclusive Project and:

- Player B chooses the **Exclusive Project** between Player B and C.

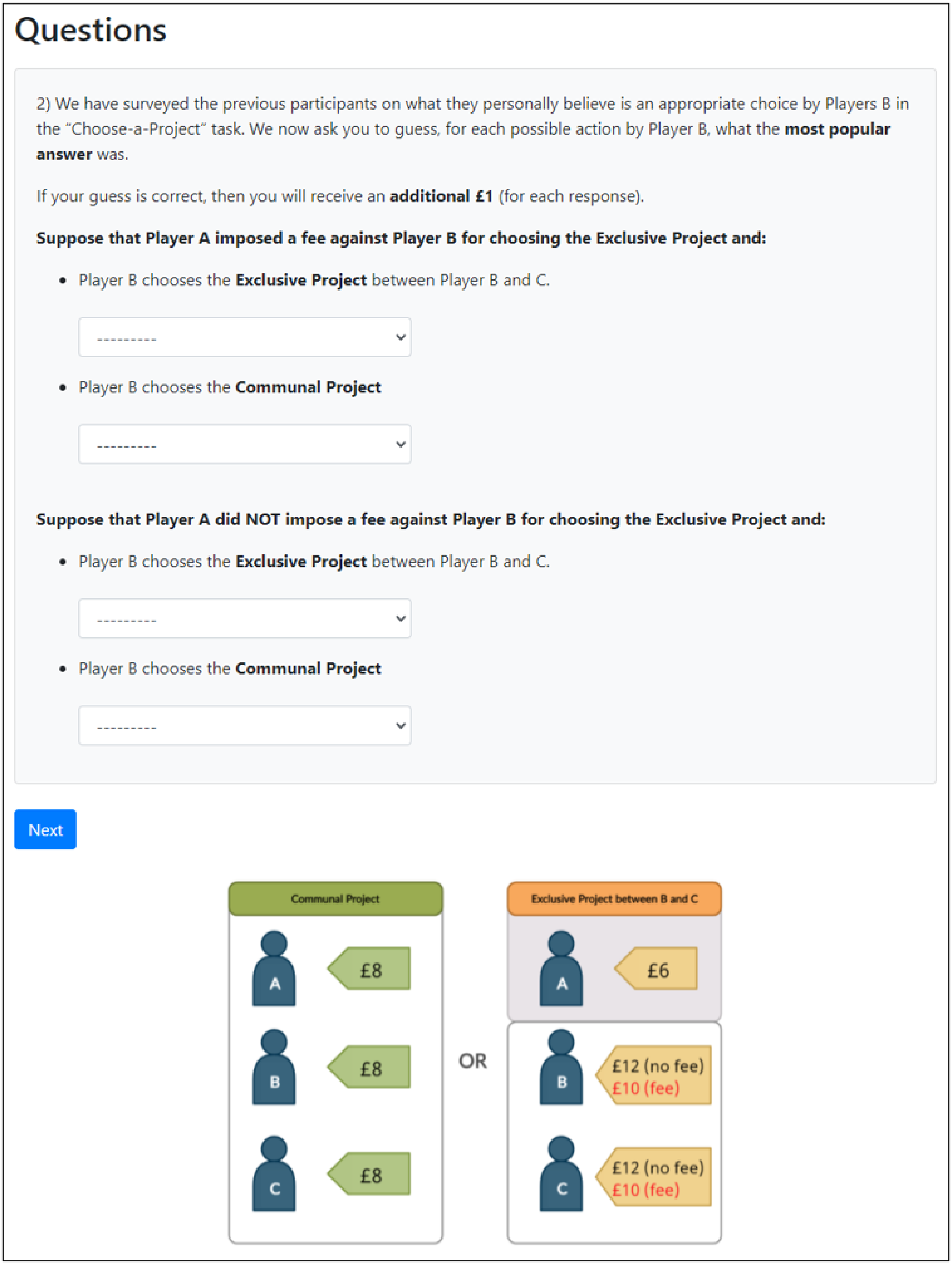
- Player B chooses the **Communal Project**

Suppose that Player A did NOT impose a fee against Player B for choosing the Exclusive Project and:

- Player B chooses the **Exclusive Project** between Player B and C.

- Player B chooses the **Communal Project**

Figure A.5: Eliciting third-party beliefs about injunctive norms (Order 1)



Questions

2) We have surveyed the previous participants on what they personally believe is an appropriate choice by Players B in the "Choose-a-Project" task. We now ask you to guess, for each possible action by Player B, what the **most popular answer** was.

If your guess is correct, then you will receive an **additional £1** (for each response).

Suppose that Player A imposed a fee against Player B for choosing the Exclusive Project and:

- Player B chooses the **Exclusive Project** between Player B and C.

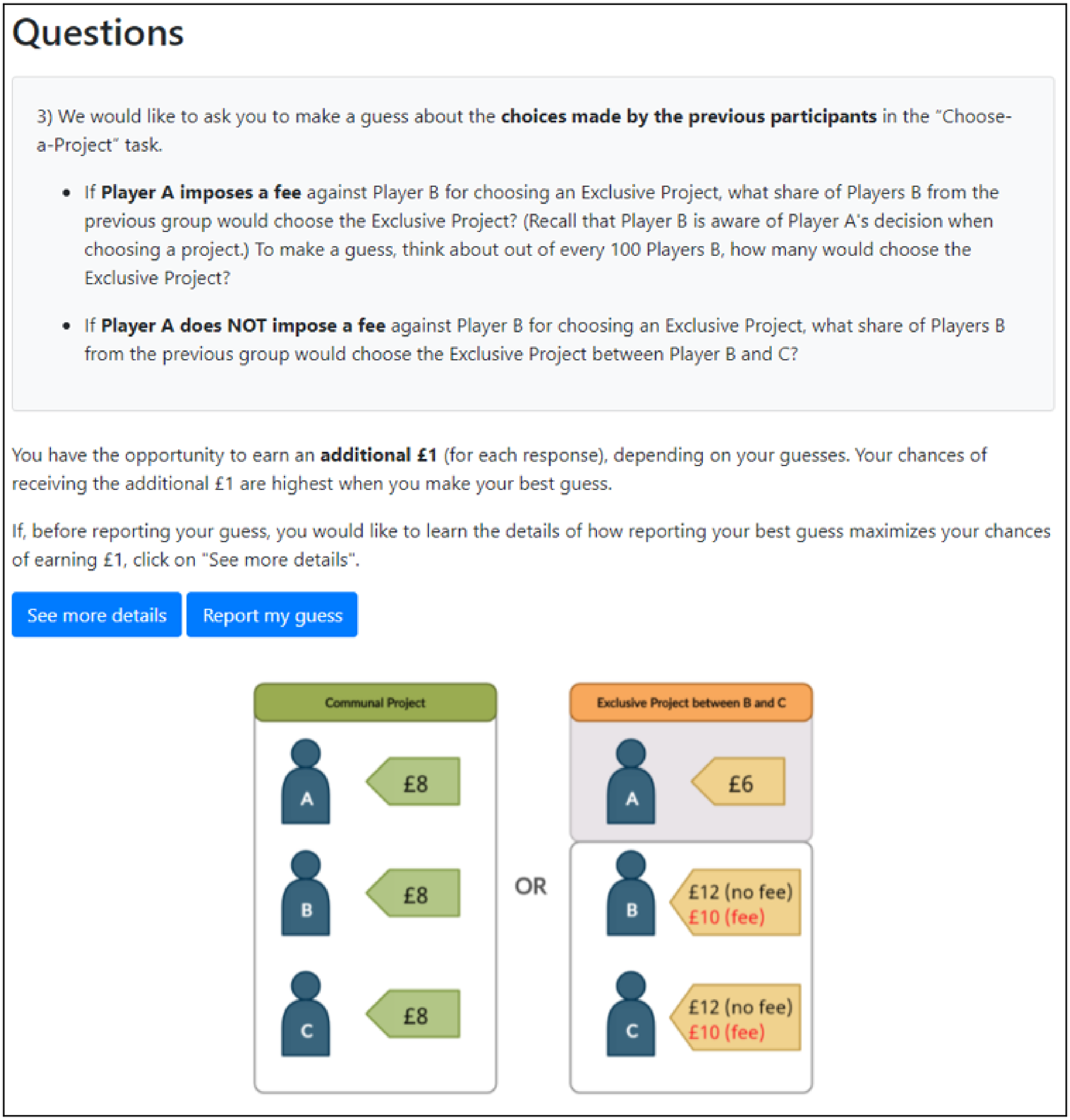
- Player B chooses the **Communal Project**

Suppose that Player A did NOT impose a fee against Player B for choosing the Exclusive Project and:

- Player B chooses the **Exclusive Project** between Player B and C.

- Player B chooses the **Communal Project**

Figure A.6: Eliciting third-party beliefs about descriptive norms (Order 1)



Questions

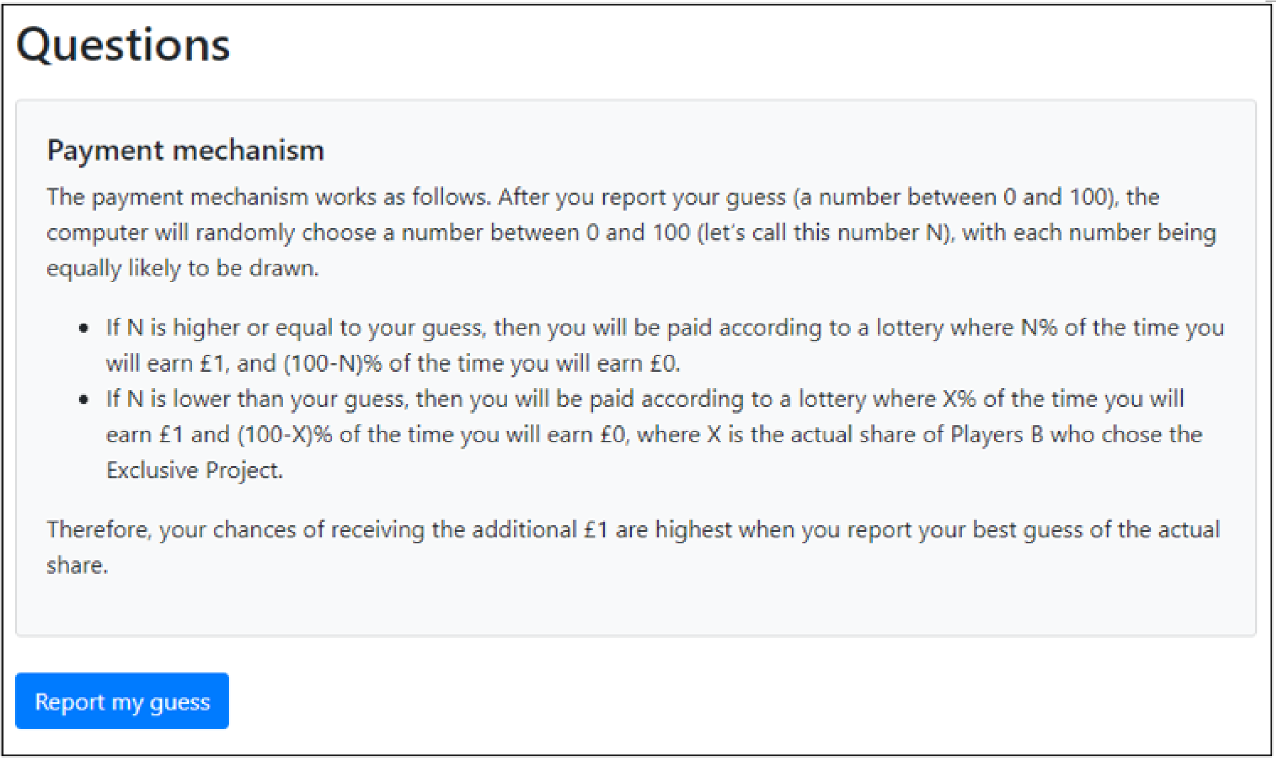
3) We would like to ask you to make a guess about the **choices made by the previous participants** in the "Choose-a-Project" task.

- If **Player A imposes a fee** against Player B for choosing an Exclusive Project, what share of Players B from the previous group would choose the Exclusive Project? (Recall that Player B is aware of Player A's decision when choosing a project.) To make a guess, think about out of every 100 Players B, how many would choose the Exclusive Project?
- If **Player A does NOT impose a fee** against Player B for choosing an Exclusive Project, what share of Players B from the previous group would choose the Exclusive Project between Player B and C?

You have the opportunity to earn an **additional £1** (for each response), depending on your guesses. Your chances of receiving the additional £1 are highest when you make your best guess.

If, before reporting your guess, you would like to learn the details of how reporting your best guess maximizes your chances of earning £1, click on "See more details".

Figure A.7: Payment mechanism



Questions

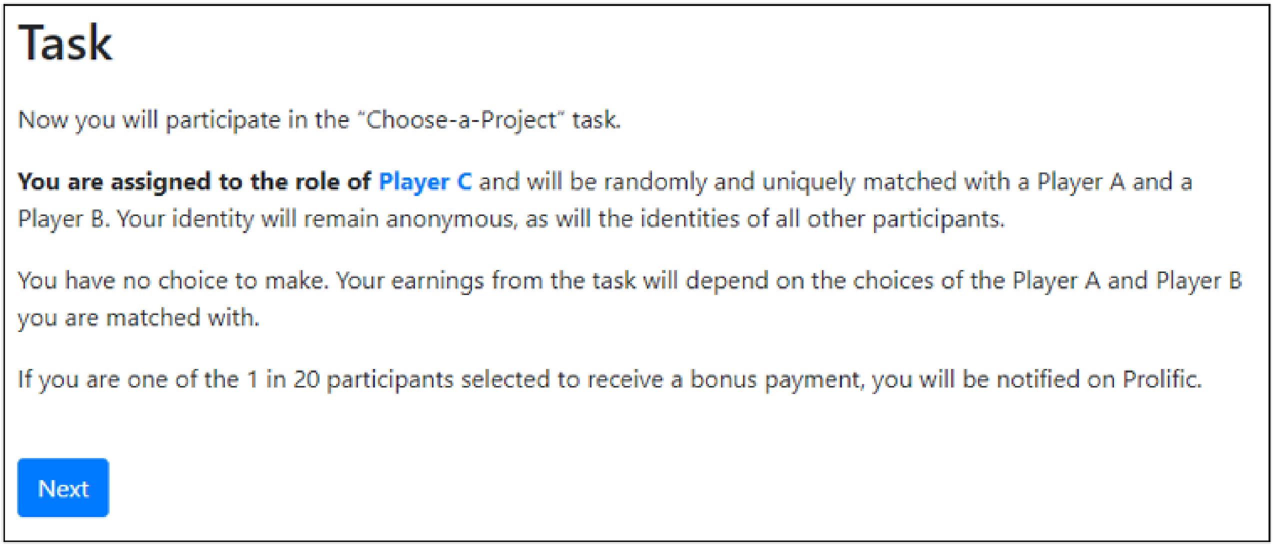
Payment mechanism

The payment mechanism works as follows. After you report your guess (a number between 0 and 100), the computer will randomly choose a number between 0 and 100 (let's call this number N), with each number being equally likely to be drawn.

- If N is higher or equal to your guess, then you will be paid according to a lottery where N% of the time you will earn £1, and (100-N)% of the time you will earn £0.
- If N is lower than your guess, then you will be paid according to a lottery where X% of the time you will earn £1 and (100-X)% of the time you will earn £0, where X is the actual share of Players B who chose the Exclusive Project.

Therefore, your chances of receiving the additional £1 are highest when you report your best guess of the actual share.

Figure A.8: The third party is informed of their role



Task

Now you will participate in the "Choose-a-Project" task.

You are assigned to the role of Player C and will be randomly and uniquely matched with a Player A and a Player B. Your identity will remain anonymous, as will the identities of all other participants.

You have no choice to make. Your earnings from the task will depend on the choices of the Player A and Player B you are matched with.

If you are one of the 1 in 20 participants selected to receive a bonus payment, you will be notified on Prolific.

B Normative beliefs

Table B.1 summarizes subjects' average personal norms (or first-order normative beliefs) while Table B.2 presents subjects' average injunctive norms (or second-order normative beliefs). In both *Self* and *Other*, across punishment and no punishment scenarios, choosing the CP is perceived to be more socially appropriate than choosing the EP ($p < 0.01$ in all comparisons, Wilcoxon signed-rank test).

Table B.1: Personal norms

	NoPun		Pun	
	CP	EP	CP	EP
Self	4.37 (0.91)	2.62 (1.19)	4.02 (1.13)	3.29 (1.03)
Other	4.36 (0.89)	2.92 (1.28)	4.24 (1.01)	2.86 (1.11)

Notes: Personal norms take a value from 1 to 5 with 1 = very inappropriate. Standard deviations in parentheses.

Table B.2: Injunctive norms

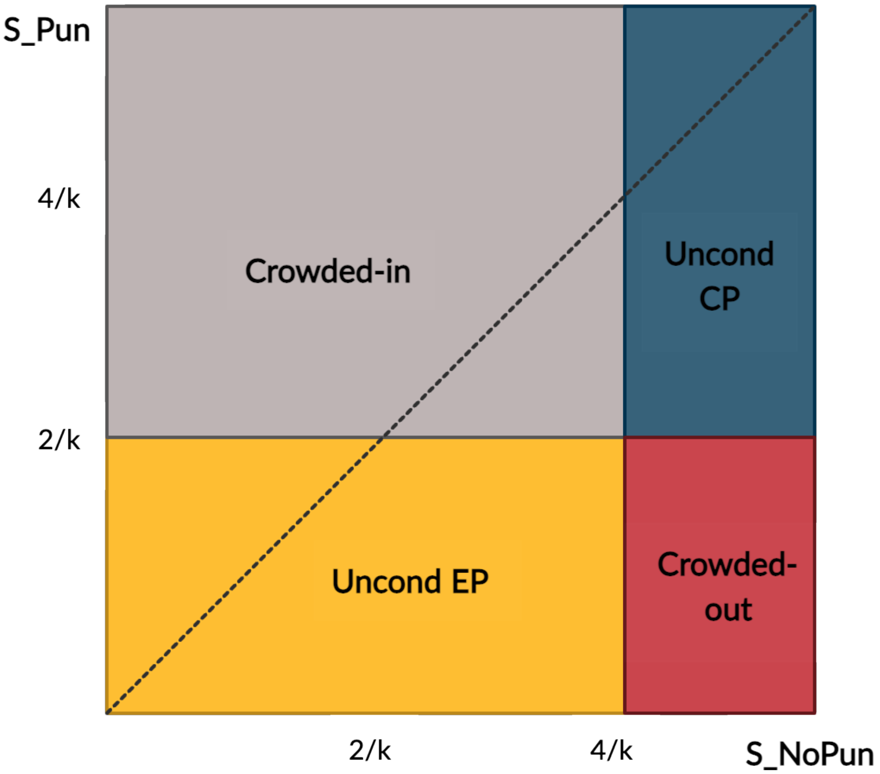
	NoPun		Pun	
	CP	EP	CP	EP
Self	4.35 (0.98)	2.59 (1.28)	3.96 (1.20)	3.34 (1.28)
Other	4.15 (1.07)	3.05 (1.34)	4.23 (1.01)	2.82 (1.24)

Notes: Injunctive norms take a value from 1 to 5 with 1 = very inappropriate. Standard deviations in parentheses.

C Agents' types

Figure C.1 presents the theoretical predictions of agents' types based on the stigma associated with choosing the EP under punishment (S_{Pun}) and no punishment (S_{NoPun}).

Figure C.1: Agents' types based on S_{Pun} and S_{NoPun}



Notes: The dotted line represents the cases in which $S_{Pun} = S_{NoPun}$, i.e. $\Delta S = 0$. The area below the line represents cases where $\Delta S < 0$, and area above the line cases where $\Delta S > 0$.

D Order effects

Table D.1 shows that the likelihood of the agent choosing the CP does not depend on the order in which the questions were asked (i.e., whether agents were first asked for their choice under punishment, or first asked for their choice under no punishment) in both *Self* ($p = 0.92$, column 2) and *Other* ($p = 0.51$, column 4).

Table D.1: Likelihood of choosing the CP

	Self		*Other*	
	(1)	(2)	(3)	(4)
Pun	1.029	1.031	1.989***	2.155***
	(0.177)	(0.194)	(0.161)	(0.178)
Order: Pun, NoPun	1.017	1.036	0.766	0.804
	(0.291)	(0.332)	(0.294)	(0.333)
Constant	0.886	0.591	0.709	1.037
	(0.213)	(1.273)	(0.224)	(1.090)
Controls	No	Yes	No	Yes
AIC	399.00	414.31	405.47	418.76
BIC	409.94	501.88	416.54	514.714
Log Likelihood	−196.50	−183.15	−199.73	−183.38
Deviance	393.00	366.31	399.47	366.764
Num. obs.	284	284	296	296

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: Odds ratio logistic regression with standard errors clustered at the individual level in parentheses. The dependent variable is the agent's choice (=1 if they chose CP). The baseline order is the choice without punishment, followed by the choice with punishment. The control variables are gender, age, education, religiosity, income and political orientation.

Concluding remarks

This thesis makes several contributions to the largely separate literature on beliefs and on prosocial behaviour. First, it offers a methodological contribution to researchers wishing to measure beliefs when there may be other motivations that compete with incentives for accuracy. Our findings highlight a challenge in eliciting beliefs and show that different elicitation mechanisms can produce different results. Second, our results provide a case for a greater diversity of perspectives in organisations as the group environment has the potential to amplify individual characteristics and lead to more biased decision making. Finally, we offer practical guidance to policymakers and practitioners aiming to design more effective punishment mechanisms that encourage prosocial behaviour: Punishment that benefits others conveys a stronger normative message and is more effective at encouraging prosociality, while punishment that benefits the punisher has greater potential to backfire.

Building on the findings of this present dissertation, an important open question is the direction of the causality between biased beliefs and social preferences, specifically: Do preferences cause individuals to bias their beliefs about others, or do biased beliefs help to justify selfish actions? One interesting avenue for future work is to investigate whether and how these biased beliefs can be mitigated in both individual and group settings, as well as the subsequent effect on behaviour. There is also scope for future research to examine other aspects of both punishment and reward mechanisms that may similarly communicate social norms and influence both beliefs about appropriate behaviour and prosocial behaviour itself.