MONASH University

MONASH BUSINESS SCHOOL

Department of Economics
ISSN number 1441-5429

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

Discussion Paper no. 2022-05

Lata Gangadharan, Philip J. Grossman and 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 nondonors 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. Addressing these biased beliefs could be more effective than trying to change underlying preferences for organisations wanting to increase prosocial behaviour.

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

JEL Classification: C9, D9, H4

Lata Gangadharan: Monash University, Department of Economics (email: Lata.Gangadharan@monash.edu); Philip J. Grossman: Monash University, Department of Economics (email: Philip.Grossman@monash.edu); Nina Xue: Monash University, Department of Economics (email: Nina.Xue@monash.edu).

monash.edu/business/economics
ABN 12 377 614 012 CRICOS Provider Number: 00008C

Belief elicitation under competing motivations: Does it matter how you ask?[*]

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

23 November, 2022

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. Addressing these biased beliefs could be more effective than trying to change underlying preferences for organisations wanting to increase prosocial behaviour.

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 either a non-incentivised (introspection), incentivised, or the incentivised Karni (2009) mechanism, a more sophisticated method that is presented 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 the benchmark is likely motivated by self-

[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 different elicitation mechanisms.

[2] The Karni 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 by Karni (2009).

image concerns. Adopting the design from Gangadharan et al. (2022), 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 beliefs, rather than by attempting to change underlying preferences. Previous studies have found

[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.

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

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 behavior. Similarly in Ging-Jehli et al. (2020), subjects who take more from another participant

[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).

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., 2022), 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 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).[10] 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. One advantage of presenting the Karni mechanism in a multiple price list format is that it allows the belief question to be structured as a question about payment (Andreoni and Sanchez, 2020). We discuss the significance of this in relation to the theoretical framework in Section 4.

We conducted additional treatments to explore whether the beliefs we elicit are robust to the timing of belief elicitation (either before or after the donation ask). We discuss these treatments in more detail in Section 6.2.

[10] Holt and Smith (2016) show that the choice menu variation of the Karni method is easier to understand and results in more accurate beliefs than the standard BDM format. Andreoni and Sanchez (2020) use a similar approach to elicit "revealed beliefs".

[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.

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. (2022) on the experimental method and validation of the method using an existing survey measure (Carpenter, 2021). This procedure by 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 several socio-demographic questions on gender, age, education, religiosity, political ideology and income. Subjects are 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* ($N =$ 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

[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).

[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%.

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

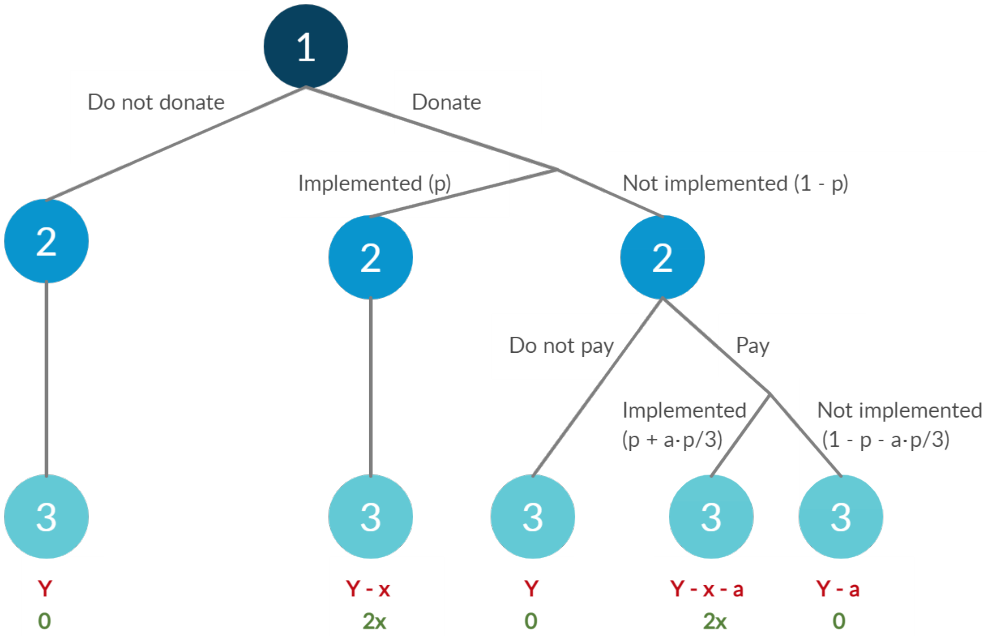
In addition to a potential financial cost of holding biased beliefs, we assume that there is also a

[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).

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 to convince a third party that the belief is accurate.[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] Based on the agent's binary donation decision, we assume that $E[\theta|X = x] > E[\theta|X = 0]$, i.e., self-image utility is higher for donors than non-

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

donors.[20] We conjecture that donors, having chosen to donate, derive sufficient self-image utility and thus have no need to bias their beliefs about others. Non-donors, however, having already decided not to give, can only protect their self-image by believing that most others in the same position also would not give. 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.

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

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

[20] 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.

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.[21] 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*.

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. (2022) 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

[21] The importance of salience in helping allocate limited cognitive resources has been studied in both psychology (e.g., Taylor and Thompson, 1982) and economics (e.g., Gabaix, 2019).

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.

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.

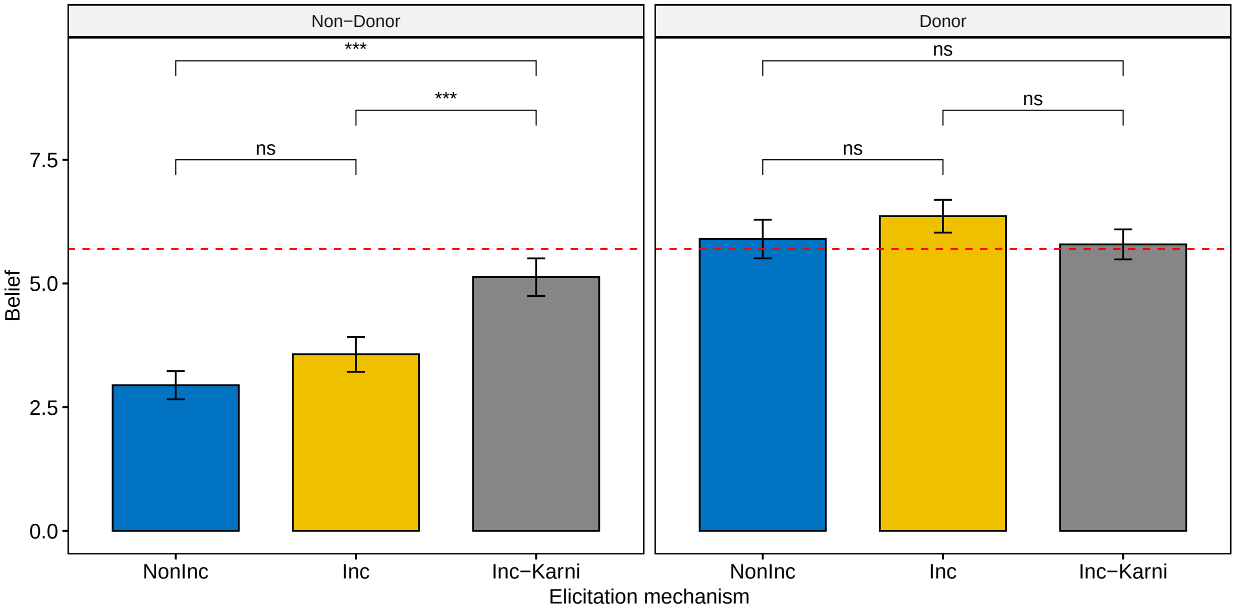
Result 1: In *NonInc*, non-donors underestimate the true proportion of donors, while donors' beliefs do not deviate significantly from the empirical benchmark.

[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.

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, $p > 0.10$).

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 $p > 0.10$.

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 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.

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.

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. (2022) 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

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

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

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.

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.

6.1 Why are beliefs different under Karni?

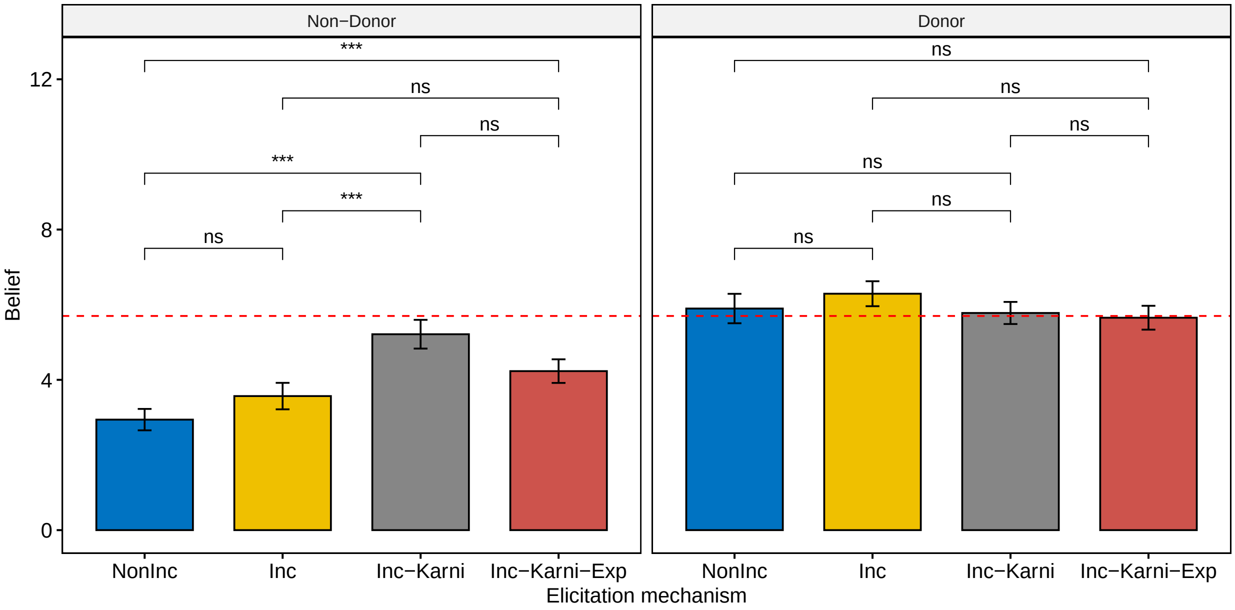
In this section we delve deeper into the mechanism driving differences between *Inc-Karni* and *Inc* and present results from an additional treatment (*Inc-Karni-Exp*) which highlight the importance of both the ability of the Karni mechanism to frame the belief question as a question about payment, and 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.

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. As a robustness check, we create a proxy for multiple switchers' beliefs by summing the number of

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

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 $p > 0.10$.

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%

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

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). A possible explanation is that the participants with inconsistent switching behaviour are also more prone to cognitive uncertainty, but these participants have already been excluded from our analysis. 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 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

[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."

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 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.

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

[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).

[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

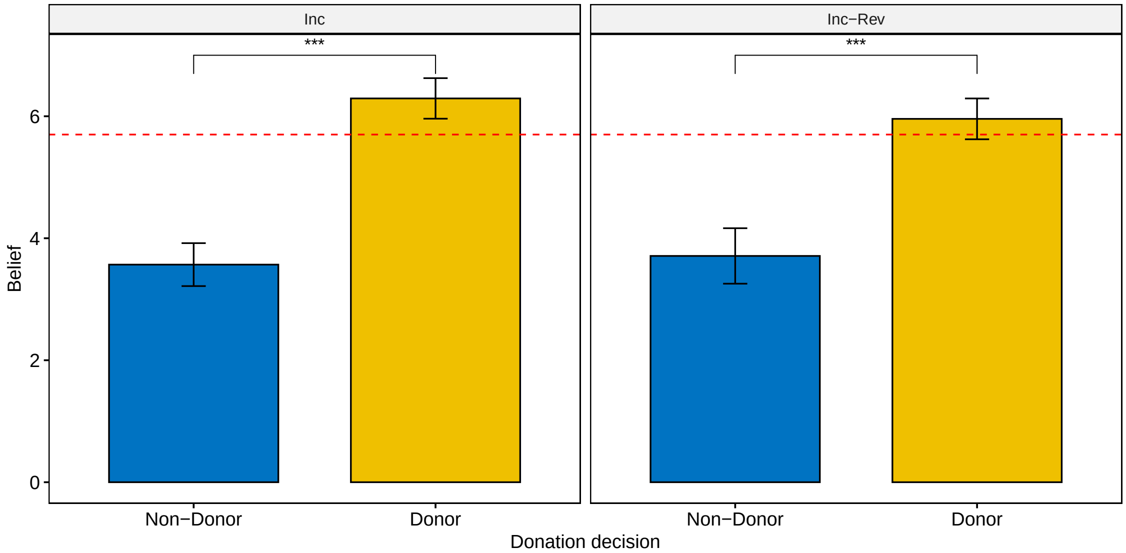
Table 5: Beliefs in the *Ask* and *NoAsk* treatments

	(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.

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

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.

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 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.

[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.

[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?"

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 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.

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.

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. (2022). Distinguishing warm glow from more altruistic giving. *Working paper.*

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.

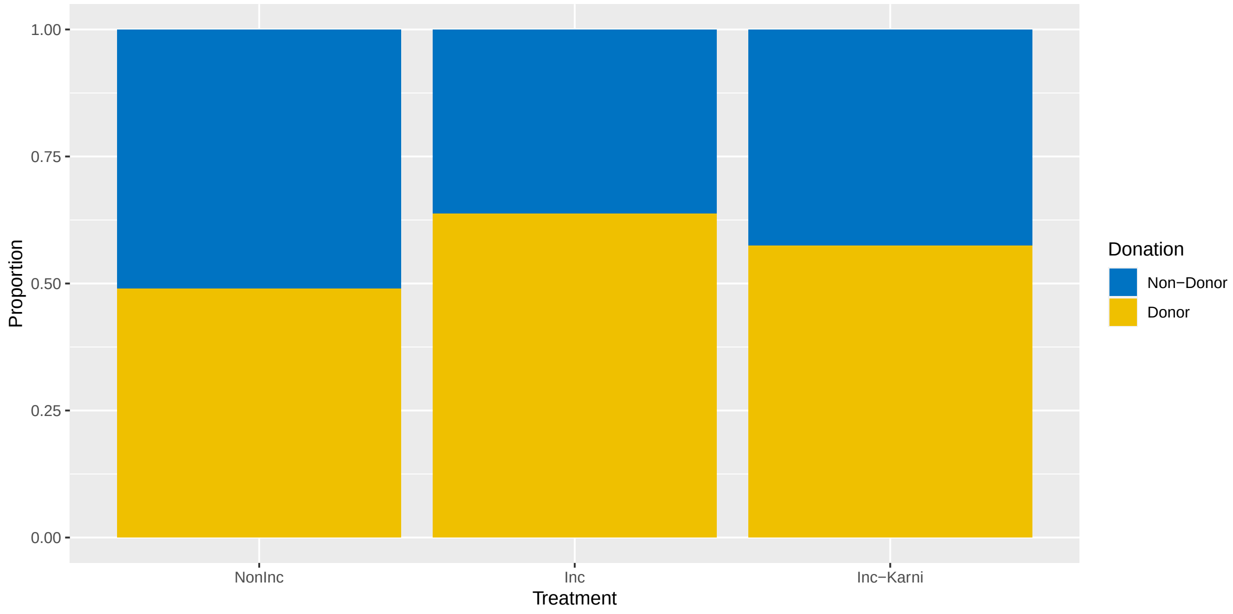
Van der Weele, J. J., Kulisa, J., Kosfeld, M., and Friebel, G. (2014). Resisting moral wiggle room: how robust is reciprocal behavior? *American Economic Journal: Microeconomics*, 6(3):256–64.

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

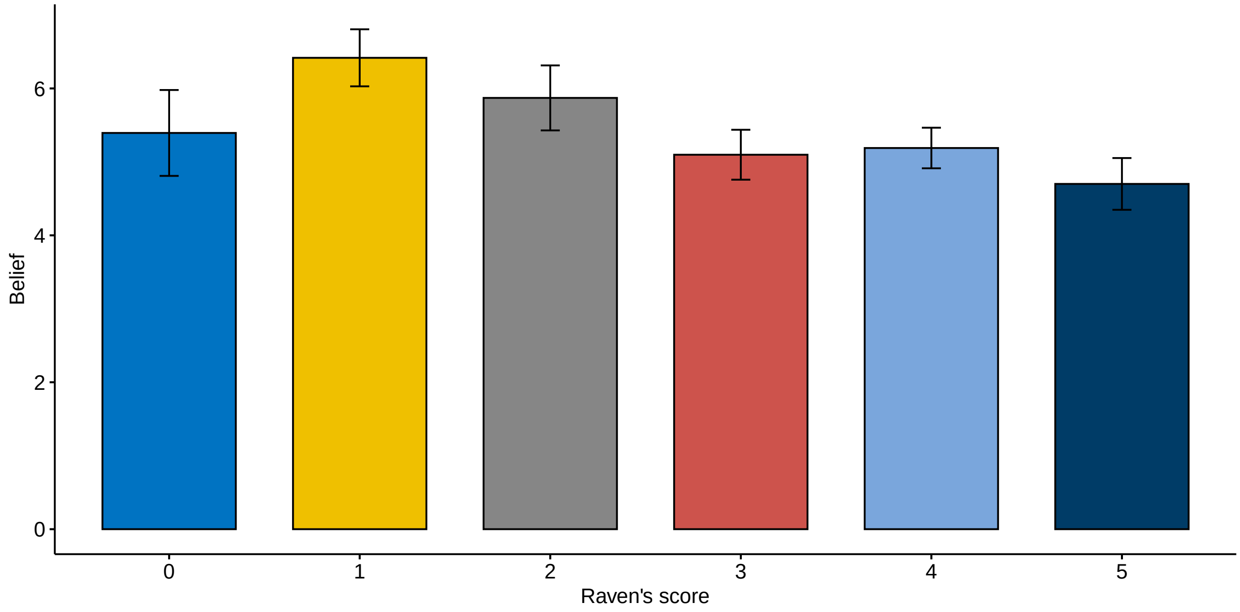
$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$

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.

D Cognitive ability

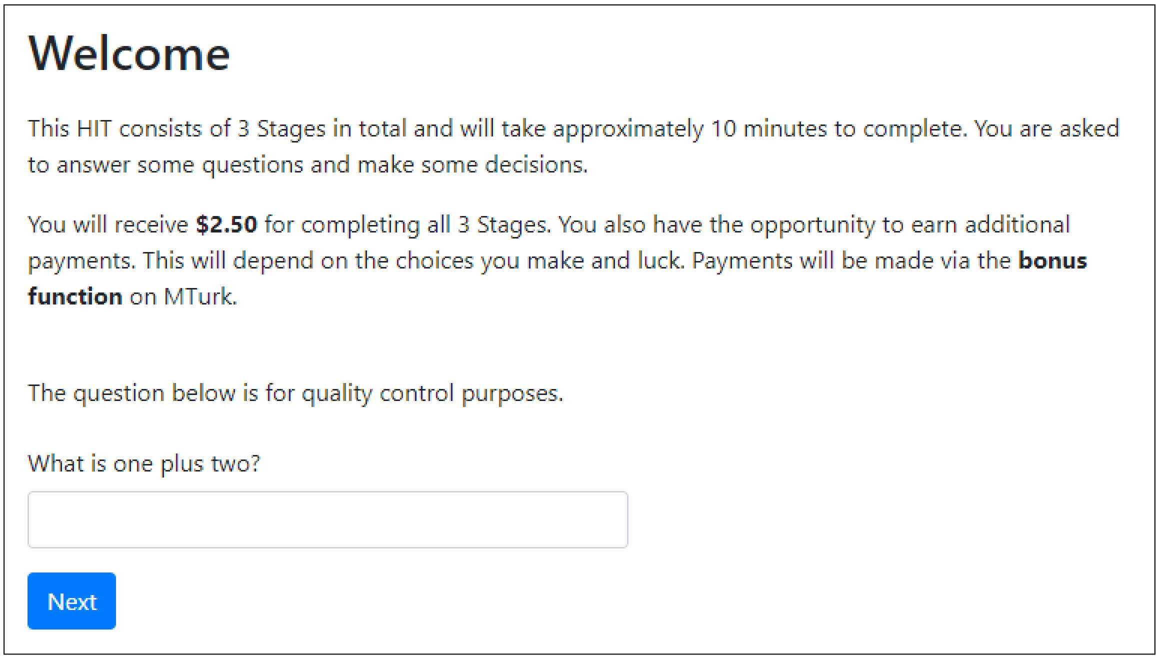
We investigate the relationship between cognitive ability and beliefs in *Inc-Karni* and contrary to Schlag and Tremewan (2021), find no clear pattern in beliefs (Figure D.1). One possible explanation for this is that we have already excluded participants who displayed inconsistent switching behaviour and therefore already exclude those who may be more prone to cognitive uncertainty.

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



Note: Error bars represent standard errors.

E Instructions



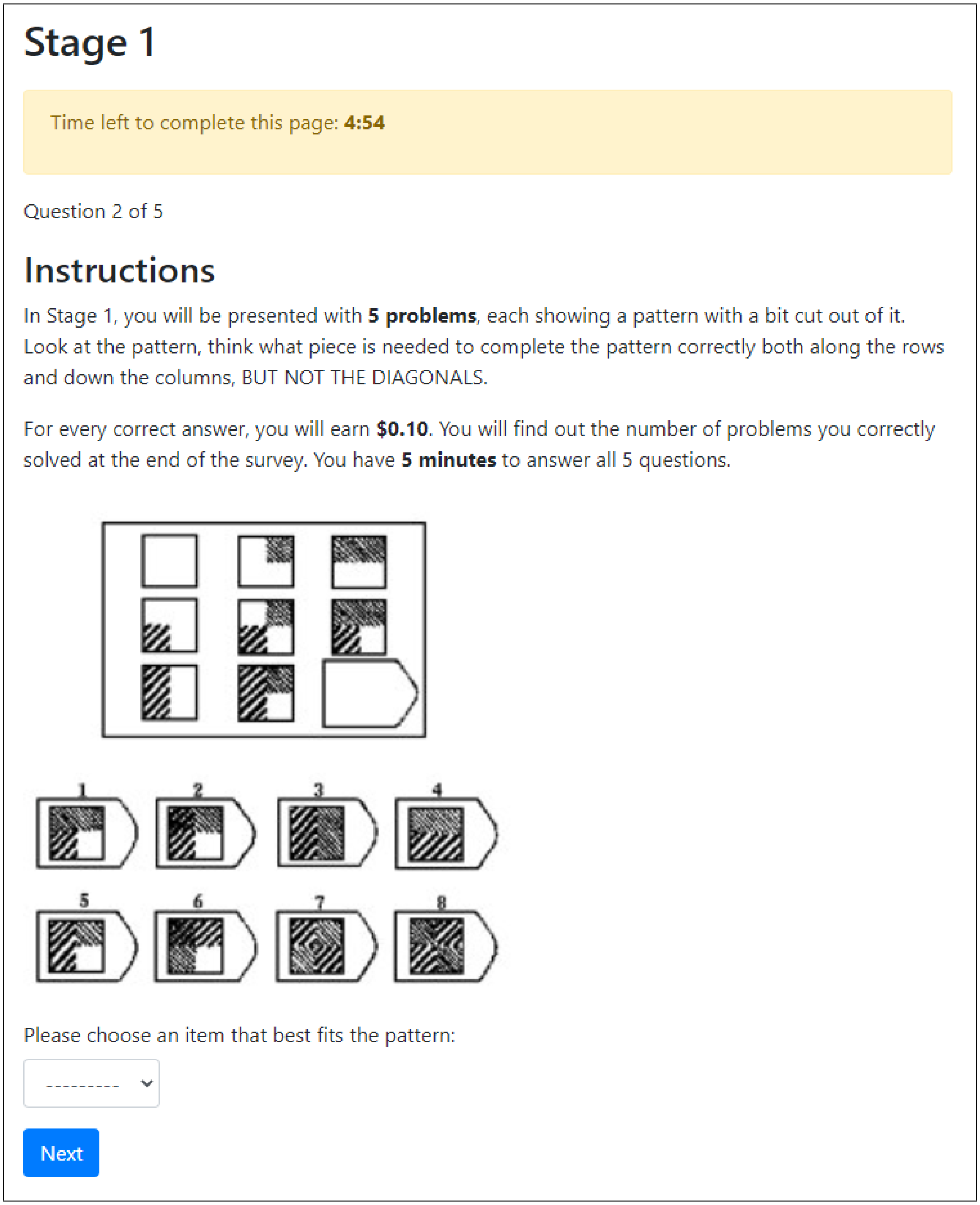
Welcome

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**

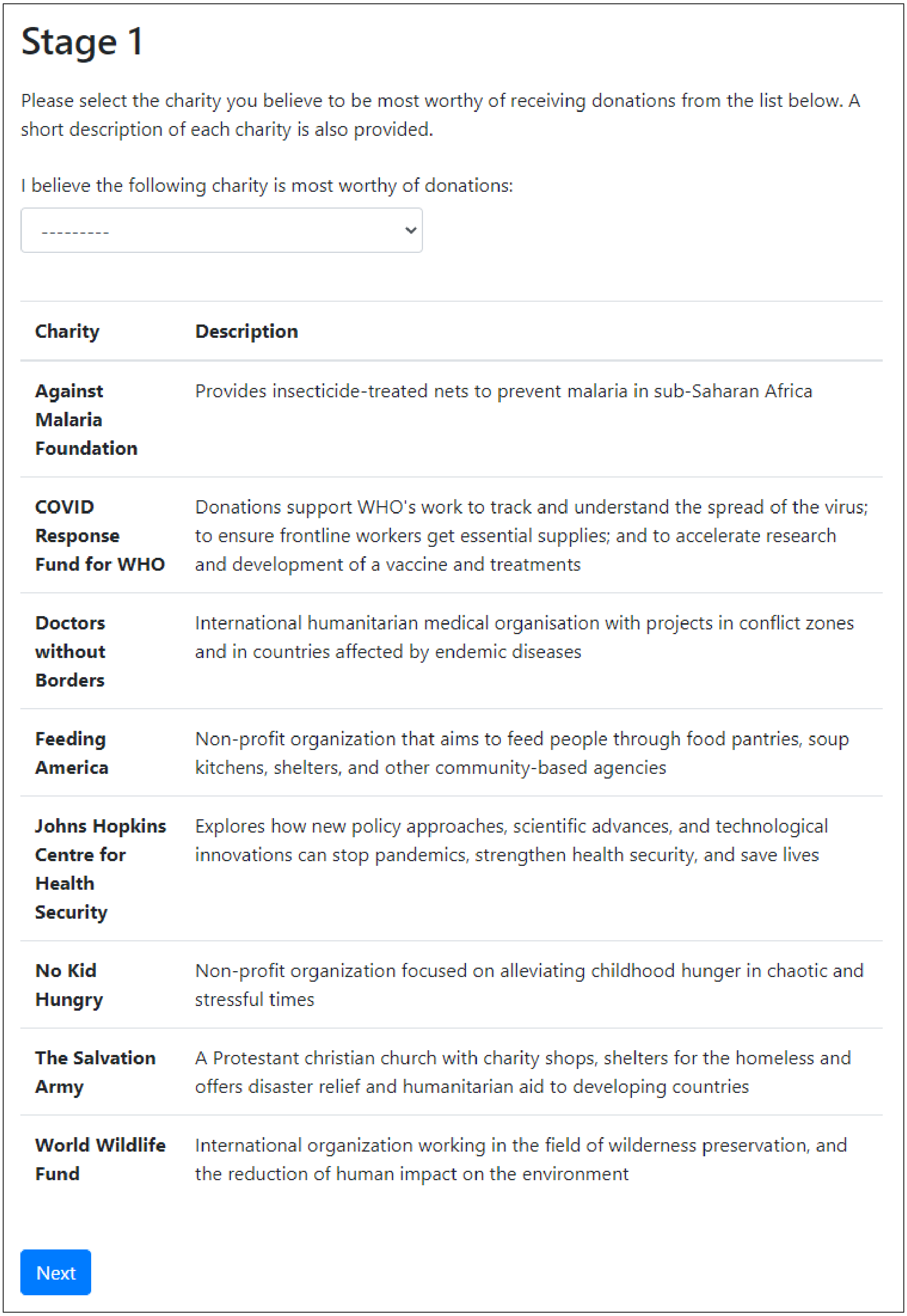
Question 2 of 5

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:

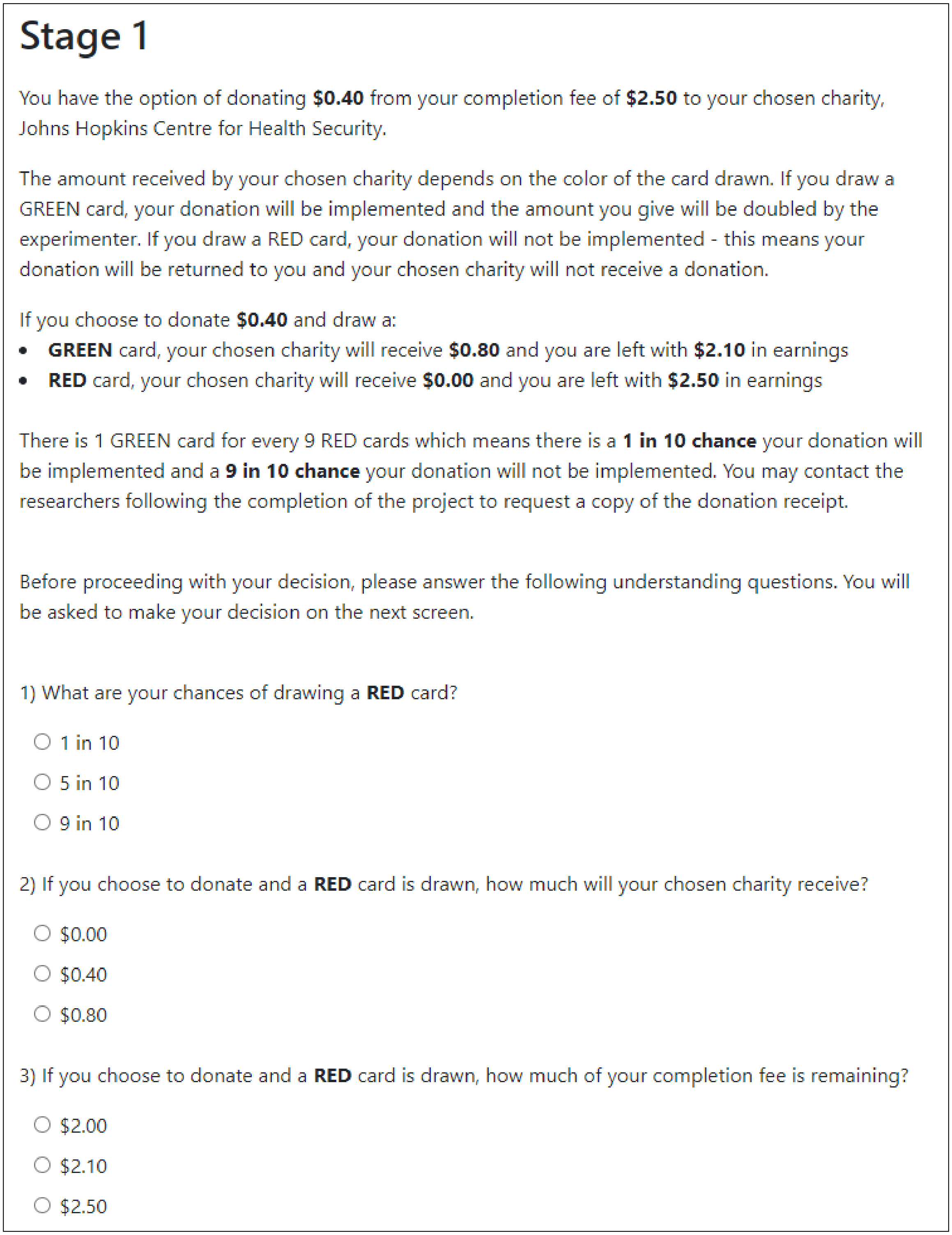


Stage 1

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

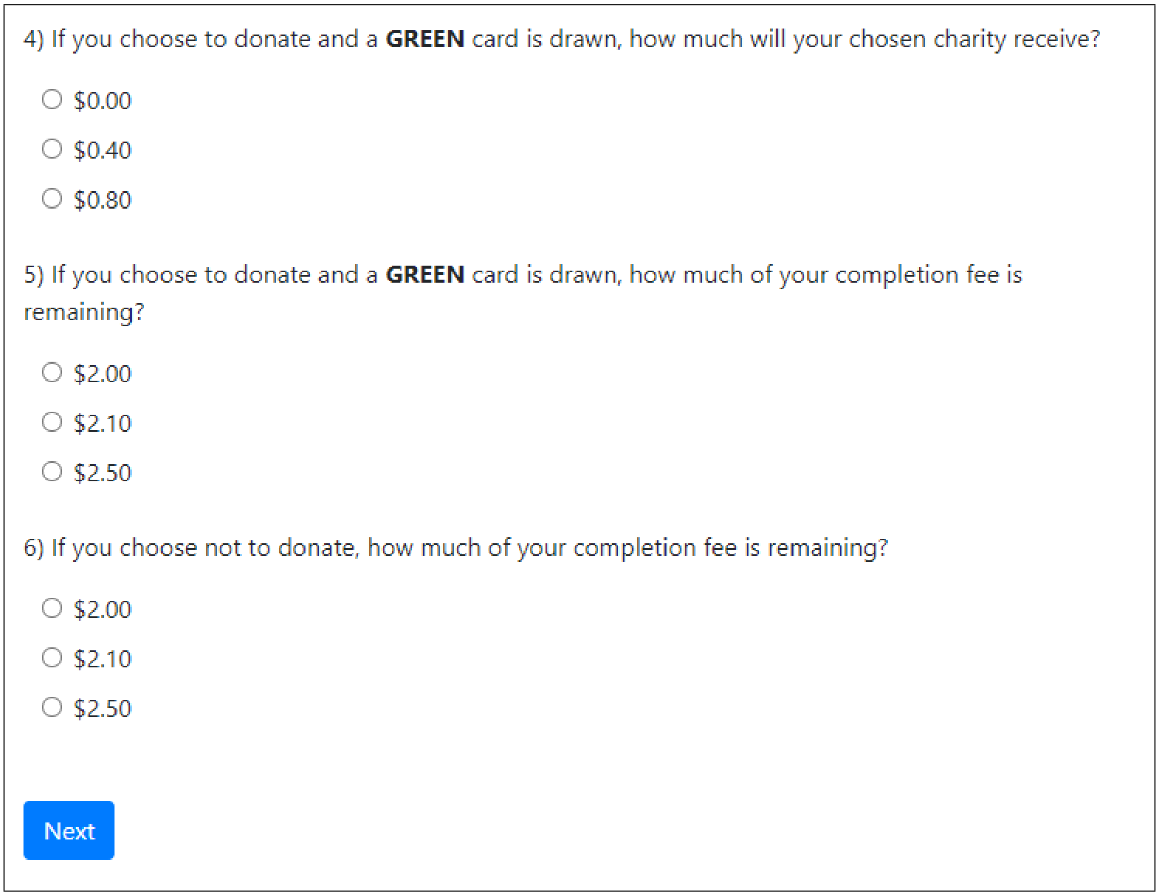
○ $0.80

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

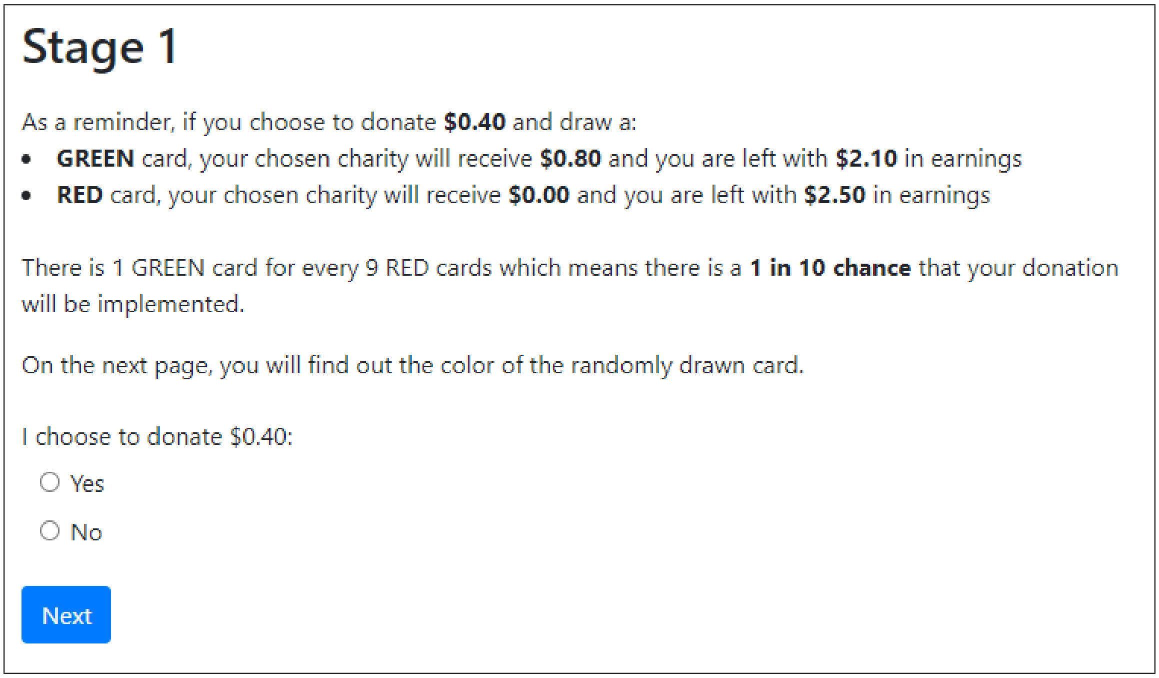
○ $2.50

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

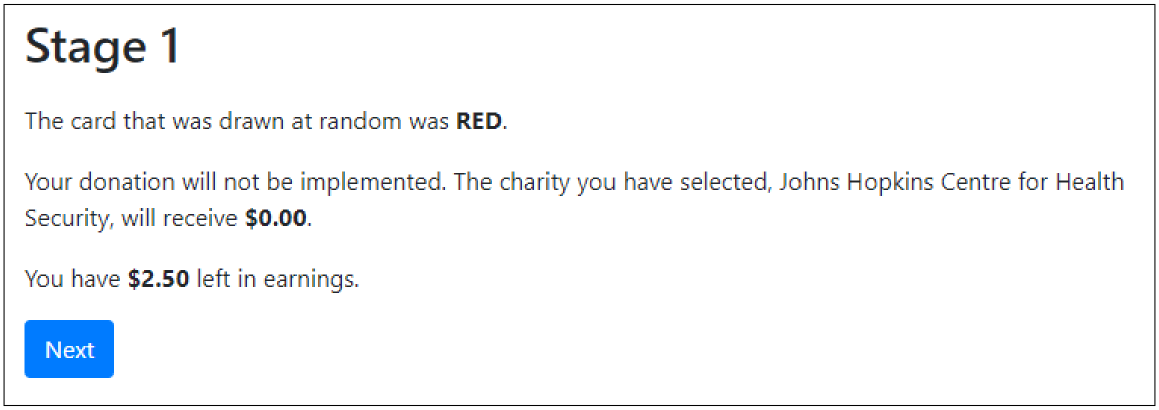
There is 1 GREEN card for every 9 RED cards which means there is a **1 in 10 chance** that your donation will be implemented.

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

I choose to donate $0.40:

○ Yes

○ No



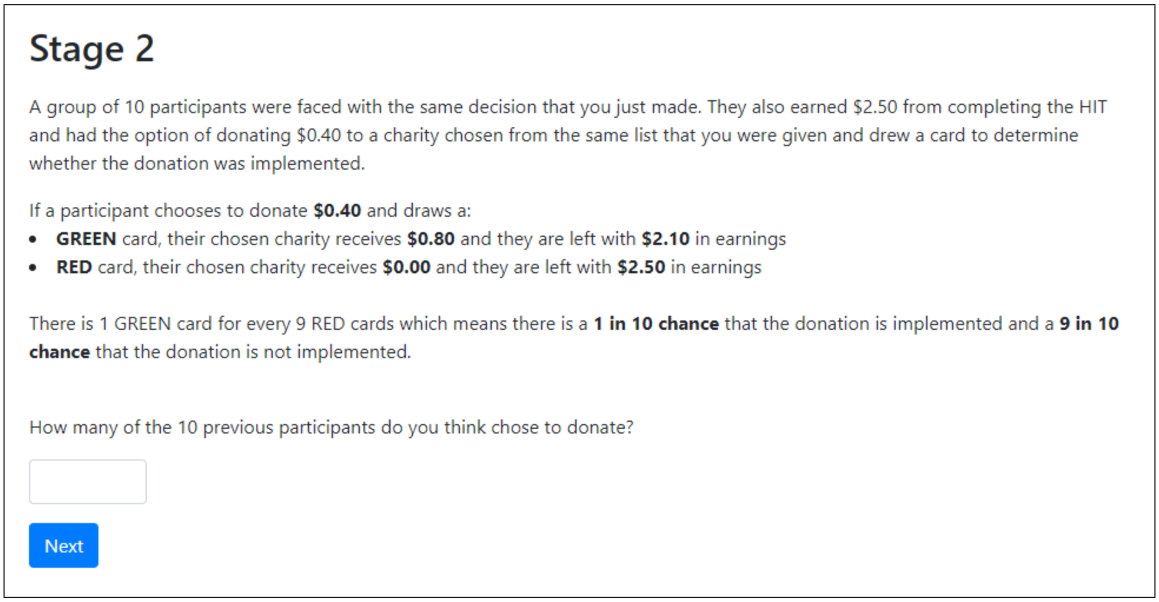
Stage 1

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.

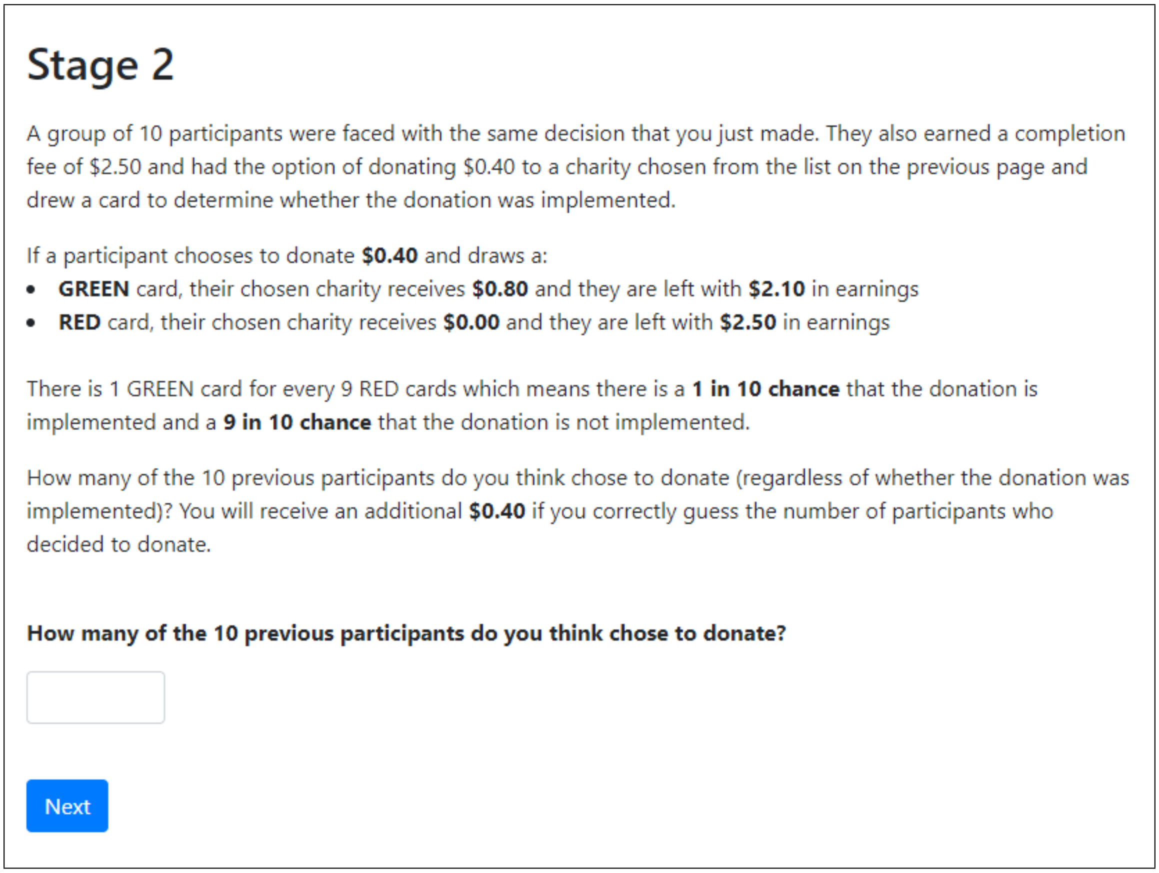
If a participant chooses to donate **$0.40** and draws a:

- **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:

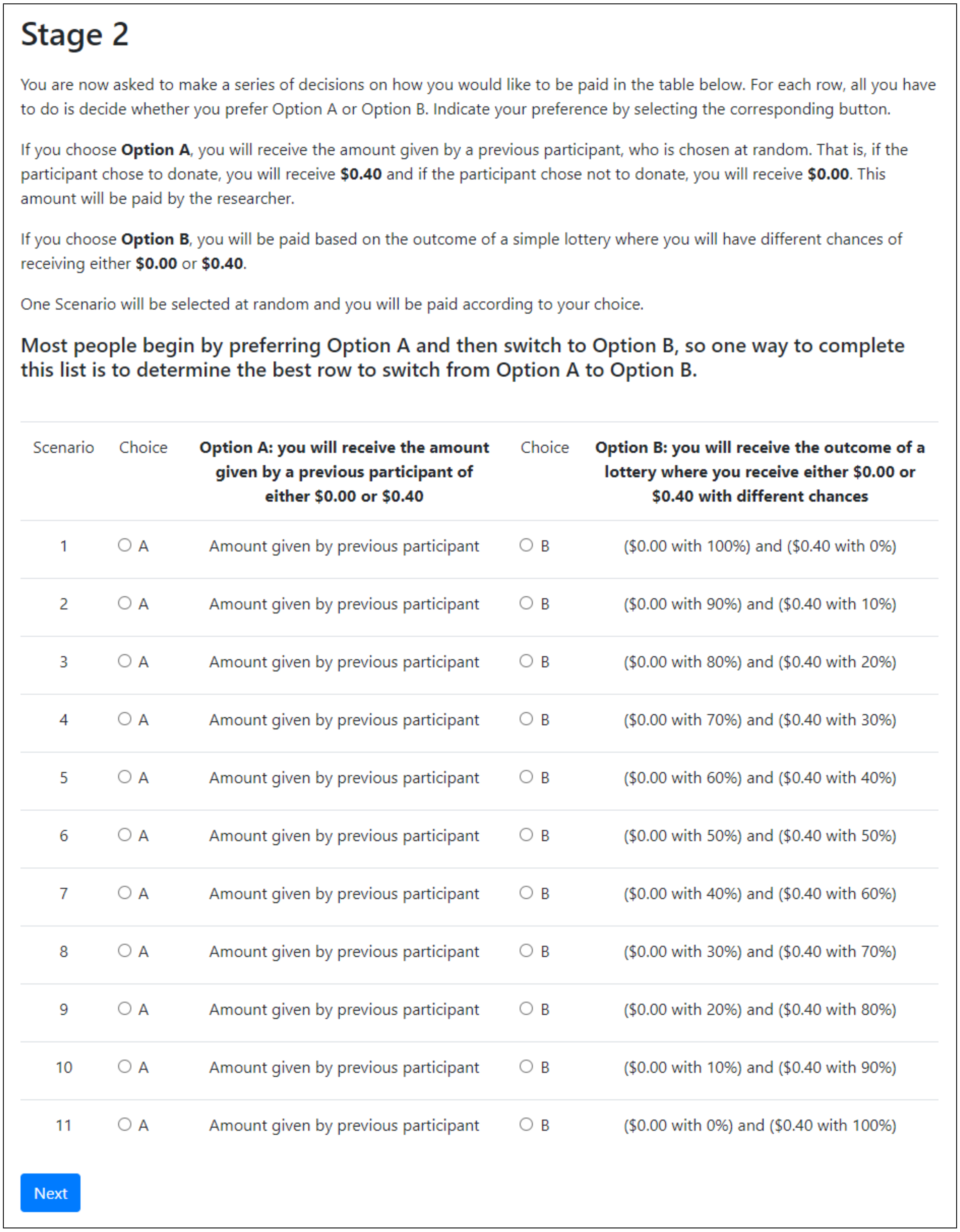
- **GREEN** card, their chosen charity receives **$0.80** and they are left with **$2.10** in earnings
- **RED** card, their chosen charity receives **$0.00** and they are left with **$2.50** in earnings

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.

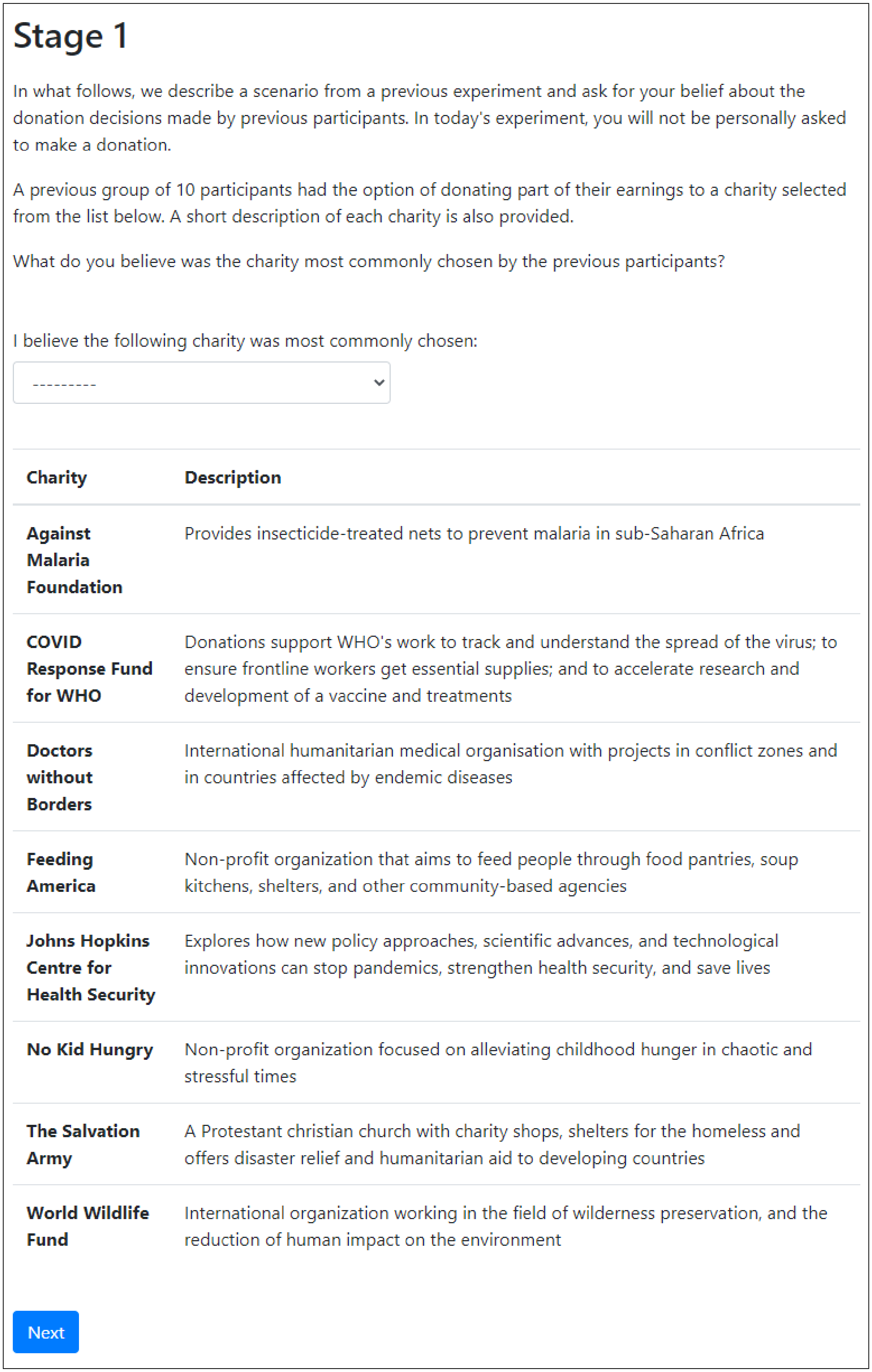
If you choose **Option B**, you will be paid based on the outcome of a simple lottery where you will have different chances of receiving either **$0.00** or **$0.40**.

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.

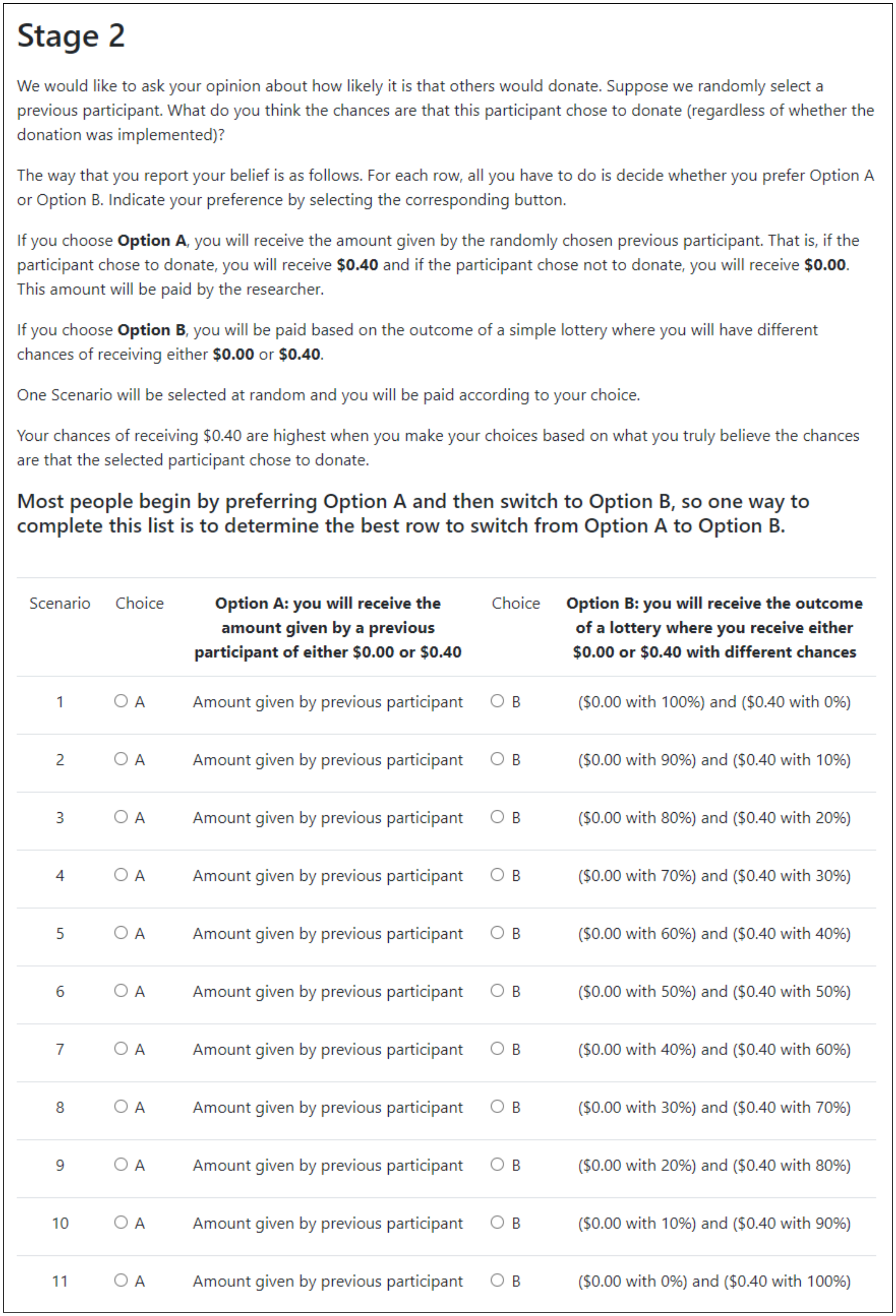
A previous group of 10 participants had the option of donating part of their earnings to a charity selected from the list below. A short description of each charity is also provided.

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%)