Cognitive Noise and Altruistic Preferences*

Niklas M. Witzig†

This version: September 2024

Abstract

I study altruistic choices through the lens of a cognitively noisy decision-maker. I introduce a theoretical framework that demonstrates how increased cognitive noise can directionally affect altruistic decisions and put its implications to the test: In a laboratory experiment, participants make a series of binary choices between taking and giving monetary payments. In the treatment, to-be-calculated sums replace plain monetary payments, increasing the cognitive difficulty of choosing. The Treatment group exhibits a lower sensitivity towards changes in payments and decides significantly more often in favor of the other person, i.e., is more altruistic. I explore the origins of this effect with Bayesian hierarchical models and a number-comparison task, mirroring the mechanics of the altruism choices absent any altruistic preference. The treatment effect is similar in this task, suggesting that a biased perception of numerical magnitudes drives treatment differences. The probabilistic models support this interpretation. A series of additional results show a negative correlation between cognitive reflection and individual measures of cognitive noise, as well as associations between altruistic choice and number comparison. Overall, these results suggest that altruistic preferences – and potentially social preferences more generally – are affected by the cognitive difficulty of their implementation.

Keywords: Cognitive Noise, Altruism, Bayesian Hierarchical Models, Experiment
JEL-Codes: C91, D91

*I thank Alexander Dzionara, Ben Grodeck, Katharina Hartinger, Marc Kaufmann, Daniel Schunk and Isabell Zipperle for helpful comments and discussion. I owe a special thanks to Dominik Straub. I gratefully acknowledge funding from the Gutenberg Academy Fellows Program and the interdisciplinary research unit IPP at Johannes Gutenberg-University of Mainz.

†Johannes Gutenberg-University Mainz, Johann-Joachim-Becher-Weg 31, 55099 Mainz, Germany

1 Introduction

Theories of social and other-regarding preferences characterize a key advancement in economics and help to explain the results of various laboratory and field outcomes irreconcilable with traditional assumptions of pure self-interest (Levine, 1998; Andreoni and Miller, 2002; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Fisman et al., 2007), with reviews in Fehr and Charness (2023); Fehr and Schmidt (2006); Cooper and Kagel (2016). Quantifying the underlying motivations of (pro-)social behavior, a growing body of research estimates population- and individual-level parameters of different social preference frameworks (e.g., Bellemare et al. 2011; Klockmann et al. 2022; Carpenter and Robbett 2022; Bruhin et al. 2019; Fisman et al. 2007 with a meta-analysis of inequality-aversion estimates available in Nunnari and Pozzi 2022).

While functional forms and parameter values differ, what these approaches share is an (implicit) assumption that social preferences are (i) a stable and fixed quantity and (ii) "fundamental", i.e., that – in a standard individual decision-making utility framework – only differences in social preferences explain differences in behavior. While the first assumption is at odds with within-person inconsistencies typically observed in experiments, the second assumption is in stark contrast with the advent of a "cognitive turn" (Enke, 2024) in behavioral economics. There, a growing body of evidence shows how cognitive imprecision, e.g., in the mental representations of objective decision problem features such as lottery payoffs and probabilities, can generate behavior like risk aversion, probability weighting and hyperbolic discounting as the result of an optimal adaptation to imprecise perceptions. In addition, this literature micro-founds inconsistencies in behavior beyond ad-hoc solution as an immediate consequence of such noisy perceptions (Woodford, 2012; Khaw et al., 2021; Woodford, 2020; Vieider, 2024b; Frydman and Jin, 2022). Similarly, the complexity of deciding according to one's preference and "cognitive uncertainty" can produce behavior previously understood as a choice anomaly and bias (Enke et al., 2023a; Oprea, 2024; Enke and Graeber, 2023).

It is natural to assume that social preferences are affected by cognitive imprecision and the complexity of their implementation as well. Tasks involving social preference require that a decision-maker assesses the (non-trivial) value of different options before deciding, rendering such operations "complex".[1] If past experiences shape social preferences, a noisy recollection of these experiences can also affect choices (see Polanía et al. (2019) for the original argument). Additionally, as social preferences are usually identified via monetary trade-offs, imprecise perceptions of numerical magnitudes – as in previous work – are a candidate, but not yet considered driver of social preference choices. First indicative evidence that noisy cognition or complexity-related processes can guide prosocial choice is provided by Enke et al. (2024). They consider the dictator game as one example how cognitive uncertainty moderates reactions to changes in objective problem features across over 30 experiments. Beyond that, empirical evidence for noisy-cognition or complexity-related effects on social preferences are still lacking, however.

In this paper, I investigate altruistic choices – a simple form of social preference decisions – through the lens of a cognitively noisy decision-maker. Based on Vieider (2024b), I develop a model of altruistic choice and show theoretically how an increase in cognitive noise can directionally affect choices. The core intuition is that higher cognitive noise

[1] Oprea (2024, 3) writes: "When we say a lottery is "complex," we mean only that its value is not transparent to the decision maker because the procedure required to optimally aggregate its disaggregated components into a value is costly or difficult."

will lead an optimal Bayesian decision-maker away from acting upon true preferences and monetary stakes and instead towards simpler mental default representations (i.e., their prior beliefs). With increased noise, the decision maker will react less strongly to changes in underlying problem features and, depending on the mental default, also choose systematically differently.

To test these implications, I implement a laboratory experiment containing two parts: In the first part, each of 300 participants makes a series of binary choices between taking a payment $self$ for themselves or giving a payment $other$ to another person. The Treatment group faces the same decision, but has the values of $self$ and $other$ replaced by to-be-calculated sums, i.e., decides between $self_1 + self_2$ $(= self)$ and $other_1 + other_2$ $(= other)$. Encasing the stakes in such to-be-calculated sums increases the cognitive difficulty of "perceiving" the monetary payments and thereby of deciding in this task. In the second part of the experiment, participants face the *identical* numerical values as previously but have to judge which of two numbers A (previously $self$), or B (previously $other$) $\times 1/2$ is numerically larger. This task mirrors the "mechanics" of the altruism decisions as participants have to compare two numbers, yet abstracts from any subjective altruistic preference with $1/2$ replacing the individual-specific and subjective altruistic-preference-dependent decision threshold with an objective and fixed term.

The main results of the experiment are as follows: In the altruism task, participants in the Treatment group exhibit (i) a flatter association between changes in payments and behavior (are less sensitive) and (ii) decide significantly more often for $other$, i.e., are more altruistic. The theoretical framework offers multiple explanations for this effect, which I begin to investigate using a probabilistic (Bayesian hierarchical) model of participants' choices. The model indicate a considerable degree of uncertainty around the mechanisms of the treatment effect, suggesting that additional data beyond the altruistic decisions is necessary to make a precise statement about the origin of the treatment effect. Herein lies the main contribution of the results in the number comparison task: In this task, although abstracting from altruistic preferences, the *treatment effect* remains qualitatively similar: The Treatment group again is less sensitive towards changes in numerical values and decides significantly less often for A (previously $self$). Interpreted together, this implies that the perception of numerical magnitudes, i.e., some intuitive prior default that $\widehat{self} < \widehat{other}$ and $\widehat{A} < \widehat{B}$, is a candidate driver for the treatment effect in both tasks. This conclusion is supported by probabilistic models based both on the number comparison data and a model estimated on a *combined* dataset, indicating a high probability for such an 'intermediate" prior belief of numerical magnitudes.

In a series of additional analyses, I further investigate associations between cognition and altruistic preferences more generally. Given identical numerical magnitudes in the altruism and number comparison task, I can closely examine potential relationships between behavior across domains: Correlation analyses show that choosing $self$ correlates with choosing A and identifying the *correct* answer in the number comparison task, suggesting that numerical cognition can play a role when eliciting altruistic preferences. Individual parameter estimates (based off of the hierarchical models) of cognitive noise correlate with performance on the Cognitive Reflection Task and Berlin Numeracy Task, showing that more cognitively-able persons are also less cognitively noisy, providing strong support for the cognitive motivation of the framework. More exploratory analyses show how measures of meta-cognition (e.g., self-reported confidence and attention) and response times – both key informative elements of choice processes – are both more strongly affected by the treatment variation and more closely related to behavior in the number comparison

versus the altruism task. This, in turn suggests that metacognitive processes could "play out" differently in domains of purely subjective preference versus domains with more objective benchmarks of choice.

With these findings, this paper predominantly speaks to three strands of literature: Primarily, the results relate to recent work on "cognitive economics". Most of this work so far focuses on the domain of risk, ambiguity, belief updating and intertemporal choice (Khaw et al., 2021; Woodford, 2020; Vieider, 2024a; Frydman and Jin, 2022; Enke et al., 2023a; Vieider, 2022). This paper shows that the core theoretical postulate, of a (Bayesian) decision maker optimally integrating noisy perceptions with prior knowledge, is applicable to the domain of (subjective) social preferences, too and offers a potential avenue for future work into the direction of cognitive noise and subjective valuations more generally. This paper also makes a methodological contribution by showing how to *causally test* the impact of increased noise beyond standard (and arguably ad-hoc) treatments of cognitive load or time pressure. The to-be-calculated sums proposed here, inspired by treatments in Enke et al. (2023a), provide an easy-to-implement method of increasing subjective uncertainty in the perception of objective problem features, that also has proven to be suitable in a more extensive repeated-trials experiment. Employing such exogenous manipulations further speaks to a broader ongoing discussion in this literature: Enke (2024, 57) outlines how it is often unclear which assumptions to make about the (locations of the) prior distribution in the Bayesian models. This paper shows how typical ignorance assumptions are not necessarily valid (see also Oprea and Vieider 2024, 33) and highlights how a combination of experimental variations increasing noise, "mirror" tasks isolating parts of the decision making process and probabilistic modelling allow to infer the parameters of prior distribution and likelihood.

This paper also relates to the literature on structural estimations of social preference parameters (Bellemare et al., 2011; Klockmann et al., 2022; Carpenter and Robbett, 2022; Bruhin et al., 2019; Fisman et al., 2007; Nunnari and Pozzi, 2022; Echeverry et al., 2023)). This paper shows how altruistic behavior (and thereby revealed altruistic preferences) can be affected by increases in the cognitive difficulty of choosing. In turn, social preference parameter estimates are thus likely to be *biased* due to the presence of unaccounted-for cognitive noise. Accordingly, classifying subjects into distinct preference types (e.g., as done in Bruhin et al. 2019; van Leeuwen and Alger 2023; Carpenter and Robbett 2022) potentially suffers from such biases. In addition, if (revealed) social preferences are used to predict or related to real-world outcomes (e.g., as in Graf et al. (2013)), the presence of cognitive noise can attenuate and bias such relationships, too. This paper makes another contribution to this literature: In an additional analysis, I show that the proposed theoretical model (of a noisy Bayesian decision maker) outperforms a standard random utility benchmark (McFadden, 1981) commonly used in this literature. The "noisy cognition" model proposed here thus offers both a theoretically more micro-founded model of altruistic choice and provides empirical arguments in its favor, motivating its application to social-preference modelling more generally.

Lastly, this papers relates to an interdisciplinary literature on dual-process models of cognition and altruism and social preferences. This literature studies differences in the level of pro-sociality between fast (more intuitive) and slow (more deliberate) decisions. For example, Rand et al. (2012) show how cooperation is largest when participants are put under time pressure, which in turn sparked a debate about whether "fairness is intuitive" (Cappelen et al., 2016) (also "social heuristics hypothesis"). The theoretical model and empirical evidence presented here add two insights to this literature: First

(i), the model demonstrates that, depending on the intuition in a given context, more intuitive (i.e., more prior-based) decision-making may also lead to more selfish choices, e.g., if monetary payments are intuitively perceived to be the same in less-for-me vs. more-for-other types of decisions. Next, while the treatment effect towards more altruism goes in a similar direction as in Rand et al. (2012), the fact that (ii), the perception of monetary payments is a likely driver for more altruism in the treatment highlights how experimental manipulations (e.g., including time pressure) may drive (pro-)social choices through channels other than via a genuine impact on social preferences per se. Hutcherson et al. (2015) put forward a comparable argument and highlight how – in light of a drift-diffusion model – individual differences in decision thresholds (which are related to decision noise) can lead to differences in (revealed) altruism independently of one's altruistic preference. Slow and fast decisions can thus be indicative of both altruistic and selfish behavior depending on the general level of altruism preference. This paper provides additional evidence in favor of this line of argument.

The remainder of this paper is structured as follows: Section 2 describes the theoretical model that illustrates how an increase in cognitive noise can directionally affect altruistic choices. Section 3 details the between-subject experimental design and differences in implementation for the Baseline and Treatment group. Section 4 introduces the results of the experiment, focusing on the main group differences in altruism and number comparison, accompanied by details on the structural estimations. Additional analyses on cognitive ability, noise and the relationship between altruism and number comparison as well as metacognition and response times follow. Section 5discusses the main results of the paper, outlining potential avenues for future research while Section 6 briefly concludes, highlighting some limitations of the current paper.

2 Theoretical Framework

The proposed theoretical model modifies models of noisy Bayesian cognition by Vieider (2024b) and Khaw et al. (2021) to accommodate for altruistic preferences.

Altruistic Preferences Imagine a decision maker (DM) who has to choose between taking a monetary payment $self$ for themselves or giving an amount $other$ to another person. They will choose $self$ if

$$(1-\beta) \times self > \beta \times other \tag{1}$$

where β is the weight the DM places on the material well-being of the other person (i..e, an altruism parameter) and its complement, $1-\beta$ is the weight the DM places on their own well-being (see e.g., Bernheim and Stark (1988); Levine (1998)). While the value of β can, in principle, be any real number, a sensible restriction will be to expect $\beta \in (0, 0.5)$, i.e., that the DM will place a positive weight on the other person's payment yet still cares more strongly about their own payment. This choice rule abstracts from many important notions relevant for social preferences, such as (dis-)advantageous inequality aversion (Fehr and Schmidt, 1999), reciprocity concerns (Falk and Fischbacher, 2006; Bellemare et al., 2011), or social norms (Carpenter and Robbett, 2022) and also does not distinguish between (non-)warm-glow giving (Andreoni, 1989). Instead, it focuses on the core trade-off akin to many types of social preference decisions: Trading off one's own vs. another person's material wealth. This rule, in turn, is similar to notions of

"pure altruism" (Levine, 1998), "preferences for giving" (Fisman et al., 2007), and "social welfare preferences" (Andreoni and Miller, 2002; Charness and Rabin, 2002) assuming a strict positive weight on the payment of the other person. Rearranging equation 1 and applying the natural log[2] to both sides gives:

$$\ln\left(\frac{self}{other}\right) > \ln\left(\frac{\beta}{1-\beta}\right) \tag{2}$$

which states that the DM assesses whether the (log) ratio of monetary payments, $\ln\frac{self}{other}$, is larger than their (log) altruism preference threshold $\ln\frac{\beta}{1-\beta}$. This structure predominantly makes the (computation of the) later model more tractable, yet expressing the payments and the preference threshold as (logs of) ratios also has a natural interpretation: $\frac{\beta}{1-\beta}$ is the weight a DM places on the other person's payment relative to their own. E.g., in the case of $\beta = 0.2$, which implies $\frac{\beta}{1-\beta} = 0.25$, the DM values each euro for the other person one-fourth as much compared to a euro for themselves. Judging monetary payments as ratios further aligns with evidence from cognitive psychology about numerical judgments (a feature I discuss below more extensively) which in turn will be relevant to many choice rules comparing monetary values.

Noisy Bayesian Decision Maker Following Vieider (2024b) and Khaw et al. (2021), I apply a Bayesian perspective to equation 2 to allow cognitive noise to affect altruistic choices based on an intuition of "perceptual uncertainty", i.e., that the perception of problem features – $\frac{self}{other}$ and $\frac{\beta}{1-\beta}$ – gives rise to a noisy mental representation of both.[3]

Noisily representing monetary payments is a feature well grounded in research from cognitive psychology: Ample evidence suggests that humans possess an "approximate number sense" for mental representations of numerosity, e.g., judging which of two boxes on a screen contains more dots (Feigenson et al., 2004). Such approximate behavior is also likely to be at play for symbolic characterizations of numbers, including that of Arabic numerals (Nieder and Dehaene, 2009; Dehaene, 2011).[4] This e.g., manifests in the "numerical ratio effect": People's performance in distinguishing between two Arabic numerals strongly depends on the numerical ratio between both numbers (Dehaene, 1993). This effect is evident in neuroimaging data and materializes during "passive viewing of numerical stimuli without an explicit behavioral task" (Cantlon et al., 2009, 2219). This suggests that the famous "Weber's-Law", which states that the necessary increase to detect a difference to a base stimulus is proportional to the base stimulus, also holds for numerical stimuli. Supporting this interpretation, studies aiming to map the "mental number line" also find evidence for a non-linear compressed mental representation of Arabic numerals (Longo and Lourenco, 2007). Furthermore, Prat-Carrabin and Woodford (2022) show that the relationship between discriminability and bias – a core law of human

[2]This follows the original model by Vieider (2024b) who demonstrates that logging the choice rule does not alter the results in a meaningful qualitative manner. See there for a derivation for the un-logged later (probabilistic) choice rule.

[3]The exact origins of this cognitive noise are beyond the scope of this paper. The general motivation for cognitive noise can be linked to the idea of a "Bayesian Brain" (Doya et al., 2006) from neuroscience, i.e., that the Brain combines uncertain sensory evidence with prior knowledge optimally.

[4]Electrophysiological recordings of monkeys can single out specific neurons favoring the mental representation of specific numbers. Crucially, the activations of these neurons are *bell-shaped*: They activate strongest at their designated neuron and less pronounced at other numbers while the activation declines in numerical difference (Diester and Nieder, 2007).

perception (Wei and Stocker, 2017) and originally formulated for sensory domains – also holds for numerical cognition.

A noisy mental representation of the preference threshold is plausible as well: Given that the true preference β remains an entirely *subjective* quantity, the DM must rely on introspection to form a belief about their preference. If past experiences shape β, an imperfect memory could introduce uncertainty around the true preference for the DM, i.e., introduce noise (see Polanía et al. (2019) for the original argument for subjective valuations).

To formalize the noisy mental representations of $\frac{self}{other}$ and $\frac{\beta}{1-\beta}$, assume that the DM obtains mental signals about the true values from a distribution of possible representations:

$$s_{\frac{self}{other}} \mid \ln \frac{self}{other} \sim \mathcal{N}\left(\ln\left(\frac{self}{other}\right), \nu^2\right), s_{\frac{\beta}{1-\beta}} \mid \ln \frac{\beta}{1-\beta} \sim \mathcal{N}\left(\ln\left(\frac{\beta}{1-\beta}\right), \nu^2\right) \quad (3)$$

where $s_{\frac{self}{other}}$ and $s_{\frac{\beta}{1-\beta}}$ are the mental signals, with both distributions sharing a common variance ν^2.[5] With the mean as a logarithm, the noise term becomes signal-dependent (as the variance of the exponentiated values increases in the mean of the original distribution), matching the intuition of Weber's Law (see also Barretto-García et al. 2023). The Bayesian DM combines the mental signals with their *prior beliefs* about both the preference threshold and the ratio of monetary payments:

$$\ln \frac{self}{other} \sim \mathcal{N}\left(\ln \mu^{\hat{r}}, \sigma^{\hat{r}^2}\right), \ln \frac{\beta}{1-\beta} \sim \mathcal{N}\left(\ln \mu^{\hat{b}}, \sigma^{\hat{b}^2}\right) \quad (4)$$

where $\ln \mu^{\hat{r}} = \ln \frac{\widehat{self}}{\widehat{other}}$ and $\ln \mu^{\hat{b}} = \ln \frac{\widehat{\beta}}{1-\widehat{\beta}}$, the default representations of the problem features (the hat indicating prior values). The values of both prior means could, in principle, take on any value, but a common assumption is that they are equal to 0, i.e., $\ln \mu^{\hat{r}} = 0 \Leftrightarrow \frac{\widehat{self}}{\widehat{other}} = 1$ and similarly $\ln \mu^{\hat{b}} = 0 \Leftrightarrow \widehat{\beta} = 0.5$. Such prior means imply that the DM intuitively does not distinguish between payments $\widehat{self} = \widehat{other}$ and treats the importance of both their own and the other person's well-being alike $1 - \widehat{\beta} = \widehat{\beta}$. This fits a possible interpretation of prior means by Gabaix (2019, 266) as "the value that spontaneously comes to mind with no thinking".

In the present setting, these intuitive priors convey a special interpretation: Assume a typical trade-off situation of $0 < \frac{self}{other} < 1$ and $0 < \beta < 0.5$. There, $\mu^{\hat{r}} = 1$ translates into a *selfish* intuition as any payment ratio of $\frac{self}{other}$ will appear *larger* with stronger reliance on the intuition, which will lead a DM to be more inclined to decide for $self$. On the contrary, $\mu^{\hat{b}} = 1$ translates into a *pro-social* intuition, as the DM intuitively cares more strongly about the other person (i.e., $\hat{\beta} > \beta$).

Differences in prior *standard deviations* also carry a particular connotation: the DM will rely more strongly upon their prior knowledge where the prior standard deviation is

[5] This assumption helps with the identification of the model within a direct tradeoff and is a common assumption in the literature (Vieider, 2024b). However, given that $s_{\frac{\beta}{1-\beta}}$ does not refer to a number, but a subjective preference, this assumption may seem less pertinent in the present setting. Note also that the previous literature applies this assumption on monetary payments and probabilities in lotteries, which both remain numerical magnitudes. I discuss a possible extension of the model in Section 4.5 relaxing the assumption of a common noise term in favor of domain-specific noise in more detail (see also Chen et al. 2023).

smaller. This implies that the DM potentially differentially relies on their prior knowledge of preferences and monetary payments.

Given likelihoods in equation 3 and priors in equation 4, the Bayesian DM arrives at the following posterior distributions:

$$\ln\left(\frac{self}{other}\right) \mid s_{\frac{self}{other}} \sim \mathcal{N}\left(\frac{\sigma^{\hat{r}^2}}{\sigma^{\hat{r}^2}+\nu^2} \times s_{\frac{self}{other}} + \frac{\nu^2}{\sigma^{\hat{r}^2}+\nu^2} \times \ln \mu^{\hat{r}}, \frac{\nu^2\sigma^{\hat{r}^2}}{\nu^2+\sigma^{\hat{r}^2}}\right)$$

$$\ln\left(\frac{\beta}{1-\beta}\right) \mid s_{\frac{\beta}{1-\beta}} \sim \mathcal{N}\left(\frac{\sigma^{\hat{b}^2}}{\sigma^{\hat{b}^2}+\nu^2} \times s_{\frac{\beta}{1-\beta}} + \frac{\nu^2}{\sigma^{\hat{b}^2}+\nu^2} \times \ln \mu^{\hat{b}}, \frac{\nu^2\sigma^{\hat{b}^2}}{\nu^2+\sigma^{\hat{b}^2}}\right)$$

with the following expected values:

$$E\left[\ln\left(\frac{self}{other}\right) \mid s_{\frac{self}{other}}\right] = \alpha \times s_{\frac{self}{other}} + (1-\alpha) \times \ln \mu^{\hat{r}}$$

$$E\left[\ln\left(\frac{\beta}{1-\beta}\right) \mid s_{\frac{\beta}{1-\beta}}\right] = \gamma \times s_{\frac{\beta}{1-\beta}} + (1-\gamma) \times \ln \mu^{\hat{b}}$$

where $\gamma = \frac{\sigma^{\hat{b}^2}}{\sigma^{\hat{b}^2}+\nu^2}$ and $\alpha = \frac{\sigma^{\hat{r}^2}}{\sigma^{\hat{r}^2}+\nu^2}$. The lower γ and α, the more the DM relies on the "intuitive" prior values, treating payments and persons alike, and the closer γ and α are to 1, the more the DM relies on the (noisy representation of the) true values of $\frac{self}{other}$ and $\frac{\beta}{1-\beta}$. Also note that ν is part of the definition of both γ and α, guiding how the DM generally weighs true evidence and prior knowledge. This is crucial for the overall impact of noise on choices, as this effect will depend on the numerical relationship between $\sigma^{\hat{b}}$ and $\sigma^{\hat{r}}$, i.e., the possibly differential reliance on both priors as well as the relationship between $\mu^{\hat{b}}$ and $\mu^{\hat{r}}$, i.e., differences in intuitions. I will discuss this point extensively below.

The expectations of the posterior distributions form the basis of the choice rule. Mirroring equation 2, the Bayesian DM will decide for $self$ if

$$E\left[\ln\left(\frac{self}{other}\right) \mid s_{\frac{self}{other}}\right] > E\left[\ln\left(\frac{\beta}{1-\beta}\right) \mid s_{\frac{\beta}{1-\beta}}\right] \tag{5}$$

and plugging in the above expressions for the posterior expectations results in

$$\alpha \times s_{\frac{self}{other}} - \gamma \times s_{\frac{\beta}{1-\beta}} > \ln \delta \tag{6}$$

where $\delta = \frac{\mu^{\hat{b}^{1-\gamma}}}{\mu^{\hat{r}^{1-\alpha}}}$. The DM will decide for $self$ if the difference of the weighted signals is larger than the weighted difference of prior means. To arrive at a probabilistic choice rule, subtract the z-score of the random variable $\alpha \times s_{\frac{self}{other}} - \gamma \times s_{\frac{\beta}{1-\beta}} \sim \mathcal{N}(\alpha \times \ln \frac{self}{other} - \gamma \times \ln \frac{\beta}{1-\beta}, \nu^2(\alpha^2+\gamma^2))$ from the equivalent z-score of equation 6, which results in the following Probit equation (see Vieider (2024b) for the original proof):

$$Pr[self \succ other] = \Phi\left(\frac{\alpha \times \ln\left(\frac{self}{other}\right) - \gamma \times \ln\left(\frac{\beta}{1-\beta}\right) - \ln(\delta)}{\nu\sqrt{\gamma^2+\alpha^2}}\right) \tag{7}$$

Choosing $self$ is thus the outcome of a *probabilistic* process in which noise ν guides both the reliance on true values versus prior knowledge (as described above) and choice variability.

The Impact of Cognitive Noise Equipped with the probabilistic choice rule, I can investigate the (numerical) impact of an increase in noise on altruistic choices. Figure 1 simulates the impact of increasing noise ν on the probability of choosing $self$ as a function of the ratio $\frac{self}{other}$ (equation 7) for varying values of $\sigma^{\hat{b}}$ and $\sigma^{\hat{r}}$. Throughout all panels, I fix β at 0.33 and $\mu^{\hat{b}} = \mu^{\hat{r}} = 1$, a common "ignorance" prior mean.

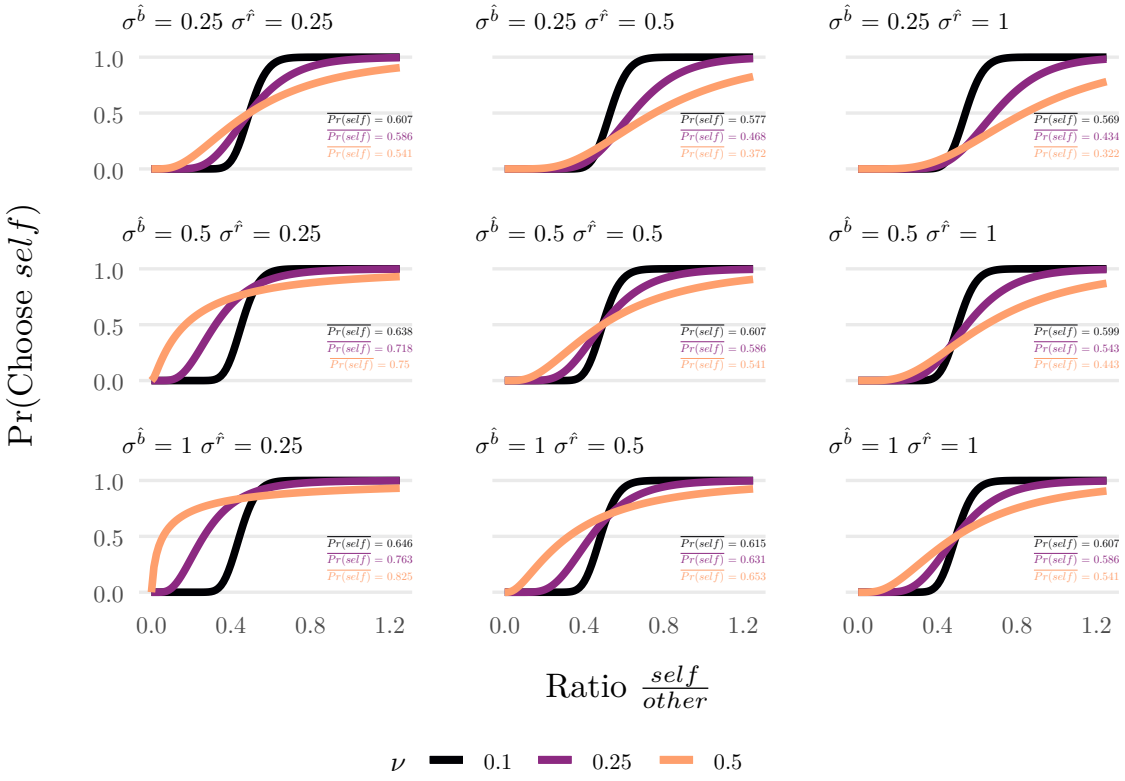


Figure 1: Impact of Noise ν on Altruistic Choices. This figure plots the impact of changes in cognitive noise ν on the probabilistic choice function (equation 7) depending on different values of $\sigma^{\hat{b}}$ and $\sigma^{\hat{r}}$. $\mu^{\hat{b}} = \mu^{\hat{r}} = 1$ and $\beta = 0.33$ in all panels. I also plot the average values of choosing $self$.

For example, consider the first column of Figure 1. Here, $\sigma^{\hat{r}}$, the standard deviation of the monetary payment prior is fixed at $\sigma^{\hat{r}} = 0.25$ and $\sigma^{\hat{b}}$ varies between $\sigma^{\hat{b}} \in (0.25, 0.5, 1)$. If $\sigma^{\hat{b}} > \sigma^{\hat{r}}$, the average probability of choosing $self$ *increases in* ν. The DM relies relatively more strongly on the payment compared to the preference prior under noise as the latter is more uncertain (i.e., $\alpha < \gamma$). The (quantitative) impact on choice behavior then is guided by the fact that with increased noise, (log) payments, $\ln \frac{self}{other} < 1$, are shrunk towards 0 more strongly compared to the preference threshold which in turn "mechanically" *increases* perceived payments on the original scale (as an attenuation of values smaller than 1 towards 0 on the log scale corresponds to an *increase* on the original scale). Note that due to the functional form of equation 7, $\ln \delta = 0$ in this case (as the prior means cancel out). However, a reformulated version of equation 7 with linear encoding and Gaussian priors[6] permits the following intuition for the impact of increased

[6] I.e., $Pr(self) = \Phi\left(\frac{\alpha \times \frac{self}{other} - \gamma \times \frac{\beta}{1-\beta} - \delta}{\nu\sqrt{\alpha^2+\gamma^2}}\right)$, where $\delta = (1-\gamma)\mu^{\hat{b}} - (1-\alpha)\mu^{\hat{r}}$

noise in this case: With $\mu^{\hat{r}} = 1$ and $\sigma^{\hat{r}} < \sigma^{\hat{b}}$, most payment ratios are perceived to be larger as they are pulled more strongly to the prior mean. In contrast, due to the higher uncertainty of the preference prior, an increase in noise does not bias choice towards the preference prior $\mu^{\hat{b}} = 1$ as much.

Increasing noise can, however, also *decrease* choices for $self$: Consider the first row of Figure 1 where $\sigma^{\hat{r}} \geq \sigma^{\hat{b}}$. Here, increasing noise shrinks $\ln \frac{\beta}{1-\beta}$ more strongly towards 0, which now translates into a *larger* preference threshold. In terms of the intuition outlined above, this parameter configuration can also be understood as a stronger reliance on the preference prior $\mu^{\hat{b}} = 1$ and thus overall *less selfish choices*.

Differences in the *mean* of the preference and payment prior can have comparable effects: Figure A.1 repeats the above exercise fixing $\sigma^{\hat{b}} = \sigma^{\hat{r}} = 0.5$ and varying $\mu^{\hat{b}}$, $\mu^{\hat{r}} \in (0.1, 0.5, 1)$. If $\mu^{\hat{b}} > \mu^{\hat{r}}$, higher levels of ν imply lower average choice probabilities for $self$. This is because $\ln \delta > 0$, i.e., a prior-based bias towards more choices for *other*. Intuitively, this can be understood as a higher level of "pro-sociality" compared to the level of "selfishness" of the payment prior. Conversely, if $\mu^{\hat{b}} < \mu^{\hat{r}}$, increasing noise increases selfish choices (as $\ln \delta < 0$).

A particular case emerges when prior standard deviations and prior means are equal, i.e., $\sigma^{\hat{b}} = \sigma^{\hat{r}}$ and $\mu^{\hat{b}} = \mu^{\hat{r}}$, depicted in the main diagonals of Figures 1 and A.1. An intuitive conjecture for this configuration is that an increase in noise simply increases randomness in behavior. However, due to the lognormal noise structure, the impact of ν on behavior is still *asymmetrical*: More probability mass will be shifted to the left of the mean, while the probability mass in the tails increases. Consequently, increasing noise will increase the probability for $self$ to the left of the indifference value (i.e., where $\frac{self}{other} = \frac{\beta}{1-\beta}$) and decrease the probability for $self$ to the right, while the latter impact is more pronounced due to the asymmetry. Overall, this translates into on average *less selfish choices*. The *indifference point* of the choice function itself, however, remains unaffected by increases in ν: Recall that, according to equation 7, the DM is indifferent between $self$ and *other* if $\alpha \times \ln\left(\frac{self}{other}\right) - \gamma \times \ln\left(\frac{\beta}{1-\beta}\right) - \ln(\delta) = 0 \Leftrightarrow \frac{self}{other} = e^{\frac{\gamma \ln \frac{\beta}{1-\beta} + \ln \delta}{\alpha}}$. If (and only if) $\sigma^{\hat{r}} = \sigma^{\hat{b}}$ and $\mu^{\hat{r}} = \mu^{\hat{b}}$, this equality reduces to $\frac{self}{other} = \frac{\beta}{1-\beta}$, i.e., remains *independent* of ν. Only differences in prior standard deviations and/or prior means shift the indifference point with increasing noise.[7]

Hypotheses Based on the theoretical model, I can also formulate hypotheses on the impact of a change in noise ν on choices for $self$. I want to emphasize that these hypotheses remain stylized in nature and should be understood as examples that do not necessarily apply to the entire parameter range (given the non-linear nature of the model), but help in organizing the mechanisms of the model nonetheless.[8] Similar to the exercise above, they should also be understood *ceteris paribus*.

Hypothesis 1_a: If $\sigma^{\hat{b}} > \sigma^{\hat{r}}$, a stronger reliance on the payment prior or (ii) $\mu^{\hat{b}} < \mu^{\hat{r}}$, a more "selfish" payment prior, an increase in ν c.p. **increases** average choices for $self$.

[7] Again referencing to the case of a reformulated equation 7 with linear encoding (and Gaussian priors), this asymmetric effect of the noise structure vanishes, whereas the remaining mechanisms of how ν affects choices described above remain intact.

[8] These hypotheses further illustrate the large degree of flexibility of the Bayesian model, depending on the assumption on the prior moments. This requires a careful interpretation of the empirical results and the exact mechanisms of a potential treatment effect in the experiment later on.

Hypothesis 1_b: If either (i) $\sigma^{\hat{b}} < \sigma^{\hat{r}}$, a stronger reliance on the preference prior or (ii) $\mu^{\hat{b}} > \mu^{\hat{r}}$, a more "pro-social" preference prior, i.e., either way, an increase in ν c.p. **decreases** average choices for $self$.

Hypothesis 1_c: If $\sigma^{\hat{b}} = \sigma^{\hat{r}}$ and $\mu^{\hat{b}} = \mu^{\hat{r}}$ (identical parameters across payment and preference prior), an increase in ν **decreases** average choices for $self$, whereas the indifference value remains unchanged.

Additional Hypotheses Outside the impact of an increase in cognitive noise emanating from the theoretical framework, other hypotheses emerge if one applies a "cognitive lens" to altruistic choices more generally.

A key assumption of the noisy cognition literature is that noisy mental representations drive choice variability and bias. Crucially, these noisy representations, e.g., of numerical magnitude, should thus have comparable effects on behavior across domains with similar "mechanics" of choice irrespective of the *subject* of the decision. Further, if "perceiving" numerical values (and subjective preferences) is person-specific, individual measures of cognitive noise should further be positively across domains within-person. Supporting evidence in this direction is presented by Frydman and Jin (2022); Frydman and Nunnari (2023), who show how lottery choice and behavior in a coordination game correlates with choices in a "perceptual" number discrimination task.

For altruistic choices, this implies that individual measures of cognitive noise and overall behavior, more generally, should be positively related to data from a comparable choice task, e.g., a number comparison task (considered in the experiment).

Hypothesis 2: There is a positive correlation between measures of cognitive noise and behavior in altruism choices and choices in a number comparison task.

Further, the noisy mental representations of problem features is generally assumed to stem from *cognitive* processes. In line with this argument, a broad class of work shows how performance in the cognitive reflection test (Frederick, 2005) – a popular tool to measure reflective thinking – empirically correlates with various biases and mistakes in choices: For instance, Augenblick et al. (2022) find that subjects who score high on the CRT infer more (less) from strong (weak) signals, Oprea (2024) finds that lower CRT performance is associated with more prospect-theoretic behavior (i.e., probability weighting and loss-aversion). Assenza et al. (2019) report a negative correlation between CRT performance and misjudgments in a portfolio valuation task and Chew et al. (2022) show a negative relationship between CRT performance and multiple switching behavior in choice list tasks. For *altruistic choices*, this implies that an an association between measures of cognitive ability and individual measures of cognitive noise ν_i – the key driver of choice bias and inconsistency – should emerge, with more cognitively higher able people exhibiting lower values of ν_i.

Hypothesis 3: Individual measures of cognitive noise negatively correlate with measures of cognitive ability.

3 Experiment

In this section, I describe the setup and implementation of the experiment, which fulfills four objectives: (i) eliciting altruistic decisions in terms of the choice rule in equation 1. (ii) Exogenously manipulating cognitive noise ν during the altruistic decisions. (iii) Eliciting choices in a number comparison task similar to the altruistic decisions, and (iv) gathering additional personal characteristics, especially regarding subjects' cognitive ability. Accordingly, the experiment consists of three parts: Part 1 entails the altruistic decisions, where cognitive noise is manipulated in a *between-subject* treatment condition. Part 2 introduces the number comparison task, and Part 3 elicits additional behavioral and survey data. All three parts are described in detail below, and a graphical overview of the experiment outline is depicted in Figure A.3.

3.1 Part 1: Altruistic Choice

Altruism In line with the theoretical setup, the main part of the experiment is centered around the following decision: taking a monetary payment $self$ (and giving nothing) or giving a monetary amount $other$ to another person (and taking nothing) as depicted in panel (a) of Figure 2. By varying the respective payments of this choice, I can infer a subject's altruistic preference. More specifically, (in the absence of noise) choices should be characterized by a unique switching point, the maximum amount of $self$ a participant is willing to forego to increase the other person's payment by $other$. I vary the monetary payments of $self$ and $other$ as follows: I chose four distinct values for $other_k$: 6.55 €, 9.26 €, 13.10 €, and 18.52 €[9] and calculate the indifference value $self_{indiff} \sim \frac{\beta}{1-\beta} \times other_k \; \forall \; \beta \in (0, 0.05, ..., 0.55)$ for all four values of $other_k$. This results in 4×12 unique combinations of $self$ and $other$ (see Figure A.4 for an illustration). Each of these combinations is repeated five times and a group of five identical trials is called a "game". Overall, subjects faced 48 games, i.e., 240 trials, in the altruism choice task of the experiment (with intermediate breaks). Following Khaw et al. (2021) I add cent values to encourage participants to approach the decisions more approximatively.[10]

At the end of the experiment, one trial is randomly drawn and implemented. Each participant is matched to a person in their session to send their chosen payment of $other$ and another person to receive the other person's choice of $other$. While the matching of the sender to the recipient is randomly determined, no participant can send to and receive from the same person and participants are instructed accordingly. Before making the 240 decisions, participants familiarize themselves with one interactive example of the choice, answer a series of comprehension questions, and encounter 12 practice trials, which are not payment-relevant and thus remain excluded from the analyses.

Treatment Condition In the between-subject treatment condition, *to-be-calculated sums* replace the monetary payments, as shown in panel (b) of Figure 2. Inspired by a

[9] Note that these values follow a series similar to the stakes in Khaw et al. (2021) as the ratio between each adjacent element in the series is a constant, i.e., $\sqrt{2}$.

[10] A critique of this approach could be that this leads participants are motivated to only focus on the main digit of the payments and simply ignore the cent values. While this is in line with an extreme form of "left-digit-bias", more recent psychological research – using eye-tracking techniques – suggests that people often pay as much attention to cents as they do to euros (Laurent and Vanhuele, 2023). Note also that e.g., Dehaene and Marques (2002, 708) explicitly avoid round numbers in their stimuli which are prices of different items.

You	Other Person
4,66 €	6,32 €

(a) Altruism Baseline

You	Other Person
3,52 € + 1,14 €	2,15 € + 4,17 €

(b) Altruism Treatment

A	B × ½
4,66	6,32

(c) Number Comparison Baseline

A	B × ½
3,52 + 1,14	2,15 + 4,17

(d) Number Comparison Treatment

Figure 2: Altruistic Choice and Number Comparison Task. **(a)** Decision screen of the Baseline condition featuring a decision between taking a payment *self* or giving a payment *other*. **(b)** Decision screen of Treatment condition, in which to-be-computed sums replace monetary values. **(c)** Baseline condition in the number comparison task. **(d)** Treatment condition number comparison. Participants always chose using the "a" ("You"/"A") and "l" key ("Other Person"/"B") on the (German) keyboard.

variation in Enke et al. (2023a), the main objective of this condition is to increase the "cognitive difficulty" of making altruistic decisions. By *disaggregating* monetary payments into two components, information processing cost will increase, which in turn may lead to mis-valuation of true incentives (see Oprea (2024) for a discussion originally about lotteries). This condition thus aims to reduce the informativeness of the mental signals, i.e., to increase ν.[11] I chose sums as relatively simple mathematical operations to allow participants to reasonably engage in the repeated trials and be able to gauge the values of the monetary payments (i.e., not reducing the informativeness too much).

The to-be-calculated sums are randomly determined but constructed in a systematic fashion: I first (uniformly) draw a random number between 0 and the smaller number of the $self$, $other$ pair. In the example, I drew $self_1 = 3.52$ € from a range between 0 and 4.66 €. $self_1$ then serves as an upper bound for a second random draw, $other_1$, i.e., 2.15 € in the example. Both determine $self_2$ and $other_2$, the complements of the sums (i.e., 1.14 € and 4.17 €). This specific procedure ensures that no matter the underlying numerical relationship between $self$ and $other$, one component of any of the two sums is larger than another component of the other sum and vice versa (e.g., in the example $self_1 > other_1$, yet $other > self$). Furthermore, the *position* of $self_1, self_2$ and $other_1, other_2$ is randomly shuffled for each participant individually. This encourages paying attention to all four components in all trials and further hinders the possibility of gauging the underlying value of $self$ or $other$ by just focusing on the positions of the components. Table A.1 provides the complete overview of all 240 trials, including the values for $self_1, self_2$ and $other_1, other_2$, which remain fixed for all participants, yet presented in random order in the experiment.

At the end of Part 1, I gather self-reported data on subjective confidence, how precisely participants calculated during the decisions, and the attention paid to both the values of $self$ and $other$ (see Figure A.13 for screenshots).

[11] An analogy to paradigms from cognitive science can also be drawn: For modeling human vision, models of Bayesian observers that integrate noisy visual perceptions with their prior beliefs have been very successful in explaining behavior. For example, experiments show that people perceive moving objects as *slower* if the contrast of the visual stimuli is low compared to stimuli with higher contrast, while the actual velocity of the object remains unchanged. This, in turn, is interpreted as evidence that people have a prior belief that things move more slowly (Stocker and Simoncelli, 2006; Weiss et al., 2002).

3.2 Part 2: Number Comparison

Number Comparison Task Part 2 of the experiment features a number comparison task. Participants have to assess which of two columns is numerically larger, either A or B $\times 1/2$ (see panel (c) in Figure 2). This task features a clear and objectively correct solution (A in the example) while aiming to mirror the "mental arithmetics" of the altruism decisions as closely as possible. The term $1/2$ replaces the threshold previously determined by each subject's β parameter (i.e., their altruistic preference) with an objective and common factor, which in turn is assumed to not give rise to a noisy mental representation, but to remain accurately perceived. This task is inspired by recent work in economics showing a correlation between elementary economic behavior and equivalent number perception (Frydman and Jin, 2022; Frydman and Nunnari, 2023).

Importantly, the values of A (B) are *identical* to those used previously for $self$ ($other$). Again, each unique combination of A,B was repeated five times. To reduce redundancy, I omit the pairs where $A = 0$ and $A > B$ in the number comparison task, such that subjects made 200 decisions in total (in 40 unique games). While the Baseline group interacts with the task as depicted in panel (c), the Treatment group again features to-be-calculated sums instead of the numerical values (d).

Similar to the number discrimination task in Frydman and Jin (2022), I incentivize this task to reward both speed and accuracy: After the end of Part 2, I calculate the share of correct solutions and the average time in seconds participants took. I then determine their earnings: 10 € × Avg. correct − Avg. time in seconds. Participants thus could earn at most 10 € if they solved every task correctly and took 0 seconds on average. Their reward was reduced for each additional second or a lower percentage of correct solutions.[12]

At the end of Part 2, I elicited beliefs about both participants' number of correct answers and the average amount of seconds they took. One of the belief elicitations was drawn randomly and determined if an additional bonus prize of 1 EUR is paid out the end according to the randomized quadratic scoring rule (Schlag and Van Der Weele, 2013; Hossain and Okui, 2013).

3.3 Part 3: Additional Data Collection

Finally, Part 3 collects several additional data from participants, which can be grouped into three different categories: (i) cognitive ability, (ii) norms and excuses, and (iii) prosociality and demographics.

Participants in the experiment have to answer six questions of the extended Cognitive Reflection Test (CRT) by Toplak et al. (2014), which entails the original three CRT questions of Frederick (2005) and adds questions similar in formulation. Figure A.14 shows a screenshot of CRT4. One of the six questions is drawn randomly and awards a bonus of 1€ if answered correctly. In addition to the CRT, I conduct the three-question Berlin Numeracy Test (Cokely et al., 2012) (unincentivized) and gather survey data on the deliberation-intuition scale (Betsch, 2004) and the German short-version of the Need

[12] I chose to implement a time-sensitive incentivization as otherwise, the task would be much more trivial to solve. Section 4.8.1 shows that the average time participants spent on the number task and the altruism decisions is identical in the baseline and even larger in the Treatment group. I read this as evidence against an argument that participants significantly decided much faster in the number comparison (which might invoke different cognition strategies) than the altruism decisions. Note also that while participants were effectively put under time pressure, there was no active reminder of their current time usage, which should help to prevent high levels of perceived time pressure.

for Cognition scale (Cacioppo and Petty, 1982) developed by Beißert et al. (2014). I chose a more extensive set of cognition-related measures to compare the standard CRT questions to alternative measures related to cognitive ability.[13]

The next block of additional data measures private and social norms and two additional survey questions about excuse-taking and (non-)altruistic behavior. I elicit social and private norms regarding behavior in the altruism task in the style of Krupka and Weber (2013), albeit in a non-incentivized way.[14] I show participants from the Baseline and Treatment group an example *both* in the Baseline and Treatment format in randomized order and ask for the subjective appropriateness of the decision (see Figure A.16 for a screenshot). I also elicit survey questions related to excuse-taking (see Figure A.15).[15]

The third block consists of several additional measures. In a simple dictator game, each participant decides how to split 10 € between themselves and another randomly determined person (see Fig A.17 for a screenshot). I instruct participants that their choice is implemented with a chance of 1%. Additionally, I obtain answers to the qualitative survey items of the Global Preference Survey (Falk et al., 2023), a visual-analog fatigue scale (Radbruch et al., 2003), as well as basic demographic information.

3.4 Implementation

The experiment ran in January 2023 at the MABELLA lab with 300 student subjects. Each subject was randomly allocated to the Baseline or the Treatment condition within an experimental session (until 150 were in each condition). As stated, subjects could earn rewards from all three parts of the experiment, and the average payment was 16.15€. The mean completion time stood at 62 minutes, while the overall session duration averaged 82 minutes, as participants had to wait until everyone in their session was finished. Instructions were presented on-screen and key screens are depicted in Appendix A.4 (translated from German). The pre-registration is available at `https://aspredicted.org/blind.php?x=5F4_72D`. The joint ethics board of Goethe University Frankfurt and JGU Mainz provided the IRB approval.

4 Results

This section presents the empirical results. This chapter first centers around differences between the Baseline and Treatment group in altruistic choices and number comparison behavior. Afterward, I discuss the additional hypotheses next to the impact of increasing cognitive noise.

4.1 Altruistic Choices: Descriptives

Figure 3 presents the average choice for $self$ for each unique value of $\frac{self}{other}$ featured in the experiment, separately drawn for the Baseline and Treatment group. The Baseline data

[13] E.g., Schunk and Betsch (2006) show that *self-reported* measures of a preference for deliberative versus intuitive reasoning correlate with individual estimates of utility function parameters.

[14] König-Kersting (2021) did not identify differences in responses between (non-)incentivizing social norm elicitation in a large-scale experiment.

[15] Based on the arguments in Dana et al. (2007) and Exley and Kessler (2024), the treatment variation could also introduce a "wiggle-room" which allows participants to make self-serving miscalculations and thereby justify more selfish behavior.

offers several insights into participants' altruistic preferences: First, perhaps unsurprisingly, the larger the payment $self$ compared to $other$, the more frequently subjects chose $self$: If $self = 0$, only 1,03 % of choices correspond to $self$, whereas, if $self > other$, 98,4 % of choices correspond to $self$. People positively care about the other person's payment, yet more strongly about their own and only very few choices are consistent with spiteful preferences. A local linear interpolation indicates that the Baseline group is indifferent (i.e., the average choice for $self$ equalling 50%) if $\frac{self}{other} = 0.474$, which implies that participants roughly care twice as much about their own payoff compared to the payoff of another person.

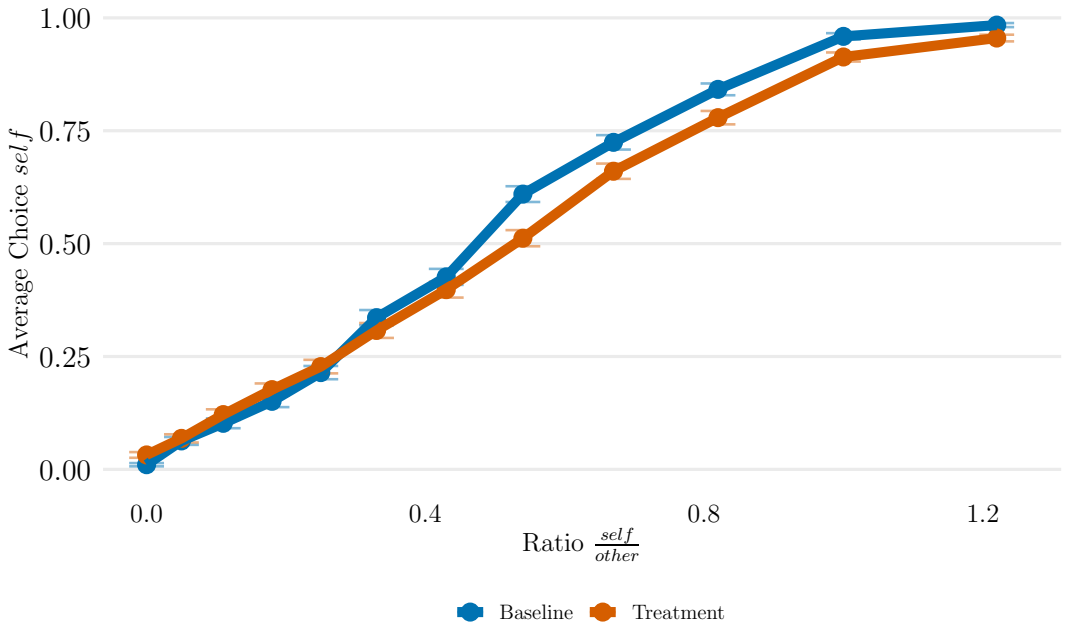


Figure 3: Altruistic choices in Baseline and Treatment group. This plot shows the association between average choice for $self$ and distinct values of the ratio $\frac{self}{other}$, separately drawn for the Baseline and Treatment Group, with 95% confidence intervals.

This is largely in line with previous evidence on (structural estimates of) social preferences, primarily that of advantageous inequality/aheadness aversion. There, the aheadness aversion parameter can be similarly interpreted as the β parameter in our framework as the weight a DM places on the well-being of another person (given the DM is better off).[16] Reviewing over 40 articles, Nunnari and Pozzi (2022) report a median value for the advantageous inequality aversion of 0.26, indicating that participants often roughly care thrice as much about their payment compared to other people's when ahead, which is in line what e.g., Bruhin et al. (2019) find. More similar to the preference of our subjects, Carpenter and Robbett (2022), Von Schenk et al. (2023) and Klockmann et al. (2022) estimate values that correspond to their subjects – as in our data – caring roughly twice as much about their own payment compared to that of another participant.

[16]Note that our framework does not distinguish between being ahead and being behind as by definition a subject is ahead if they choose $self$ and behind if they choose $other$. Thus, the present setup does not allow to separate these two motivations but comprises them into one. The fact that subjects overall substantially weigh the other person's payment could also be related how the the decision in Figure 2 is displayed which does not include the 0 € consequence for either person.

In the Treatment condition, the statements about the altruistic preferences remain largely true, albeit with subtle differences: First, the association between the average choice for $self$ and changes in the underlying ratio of $\frac{self}{other}$ is *flatter* compared to the Baseline condition. For small values of $\frac{self}{other}$, the Treatment group decides more often for $self$, e.g., 3,2 % of choices correspond to $self$ if $self = 0$, yet less often for larger values of $\frac{self}{other}$ as only 95,53 % of choices correspond to $self$ if $self > other$. Over the entire set of trials, the Treatment group behaves less selfishly: While the Baseline group decided in 45,18% of choices for $self$, the Treatment group chose $self$ in 42,93 % of the cases. This difference is statistically significant, as indicated by a two-sided t and a Fisher exact test (both $p < 0.001$). The ratio of $\frac{self}{other}$ required for indifference of the Treatment group corresponds to 0.528, a 5.4 percentage points larger ratio compared to the Baseline. Using a linear probability model, Table A.2 confirms that both the overall level of $self$-choices is 2.2 percentage points lower and that an increase in $\frac{self}{other}$ by 1 has a 6.7 percentage points lower effect on choices for $self$ in the Treatment group (see also a Probit model in Table A.3). Although their underlying preference arguably remains the same as in the Baseline, the Treatment group thus shows a dampened reaction to changes in the incentives and chooses $self$ significantly less frequently, i.e., behaves more altruistic.

Result 1: *The Treatment group shows both a flatter association between changes in payments and choices and is more altruistic compared to the Baseline group.*

Both a flatter association between varying payments and choices and a *bias* towards more altruistic choices can be rationalized with an increase of noise ν.[17] Recall Figures 1 and A.1, which outline the impact of noise on the probability of choosing $self$ (equation 7). The described treatment effect can originate from various parameter configurations: Either if (i) participants rely more strongly on a "pro-social" preference prior compared to the "selfish" payment prior (i.e., $\sigma^{\hat{r}} > \sigma^{\hat{b}}$ and $\mu^{\hat{b}} = \widehat{\frac{\beta}{1-\beta}} = 1 = \mu^{\hat{r}} = \widehat{\frac{self}{other}}$) ($H_{1a}$) or (ii) an intuition that the values of the monetary payments are unequal (i.e., $0 < \mu^{\hat{r}} = \widehat{\frac{self}{other}} < 1$), biasing the "perception" of numerical magnitudes towards some intermediate value (H_{1b}). A third possibility (iii) is a "pure noise" effect due to the lognormal noise structure (H_{1c}).

4.2 Altruistic Choices: Probabilistic Model

Given the high flexibility of the theoretical model and the corresponding multiple potential mechanisms of the treatment effect, I turn to probabilistic modeling to estimate the (posterior) probability of the parameter values given the experimental data. I use Bayesian estimation techniques, which are gaining popularity in experimental economics (see Bland (2023) for an overview and the tutorial by Vieider (2024a)). The main reason to use Bayesian techniques lies in their practicality: Because they are more flexible than, e.g., maximum likelihood estimation, they can deal more easily with more complex models and still produce meaningful uncertainty estimates of the parameters of the model (Gelman et al., 2021, 4). Here, I estimate a *Bayesian Hierarchical Model*, that determines the prior for the individual parameter values from the data. In hierarchical models, indi-

[17] Section A.3.3 discusses if the chosen treatment variation might have invoked behavior other than an increase in cognitive noise. The difference between Treatment and Baseline can not be explained by (i) an exclusive focus on (and comparison of) the first component of the sums (Figure A.10) or (ii) that the treatment only works for larger numbers (Figure A.10 and Table A.18).

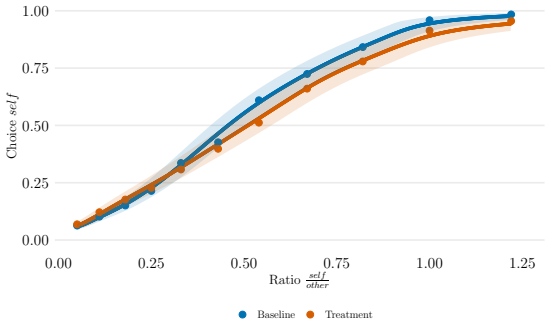
vidual parameter estimates are partially pooled towards the group mean, which reduces overfitting and thus increases out-of-sample performance (Kruschke, 2015). Furthermore, the hierarchical setup allows us to represent potential treatment differences in specific parameters efficiently by allowing (some) hyper-parameters to differ between conditions c. More specifically, the hierarchical model assumes that the individual parameter vector $\boldsymbol{\theta_i} = \left(\nu_i, \beta_i, \mu_i^{\hat{r}}, \sigma_i^{\hat{r}}, \mu_i^{\hat{b}}, \sigma_i^{\hat{b}}\right)$ of individual i is – on the log-scale – drawn from a multivariate normal distribution:

$$\boldsymbol{\theta_i} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{8}$$

where $\boldsymbol{\mu} = (\mu_c^{\nu}, \mu^{\beta}, \mu^{\mu^{\hat{r}}}, \mu^{\sigma^{\hat{r}}}, \mu^{\mu^{\hat{b}}}, \mu^{\sigma^{\hat{b}}})$ is the vector of the population-means of the parameter distributions. Note that μ_c^{ν} is allowed to differ between Baseline and Treatment group (μ_B^{ν}, μ_T^{ν}), whereas all other hyper-parameters remain identical across conditions. $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\tau})\Omega\text{diag}(\boldsymbol{\tau})$, where Ω is the correlation matrix of individual parameters and $\boldsymbol{\tau}$ is a vector of standard deviations. The hierarchical model requires to specify prior distributions for all hyper-parameters and I choose so-called weakly informative priors (see Section A.3.1 for details and prior predictive checks, also see Gelman et al. 2015 and Gelman 2020). I estimate the model with Numpyro (Bingham et al., 2019; Phan et al., 2019).

	mean	median	sd	hdi 2.5%	hdi 97.5%	$\hat{R}$
Base Parameters:						
Altr. Preference β	0.338	0.338	0.018	0.304	0.373	1.01
Prior Std. Outcomes $\sigma^{\hat{r}}$	0.93	0.86	0.326	0.495	1.536	1.01
Prior Mean Outcomes $\mu^{\hat{r}}$	0.964	0.897	1.422	0.503	1.492	1.01
Prior Std. Preference $\sigma^{\hat{b}}$	0.716	0.646	0.337	0.376	1.229	1.01
Prior Mean Altr. Preference $\mu^{\hat{b}}$	0.816	0.687	0.649	0.299	1.56	1
Group Specific:						
Noise Baseline ν_B	0.277	0.276	0.024	0.232	0.326	1
Noise Treatment ν_T	0.338	0.337	0.031	0.281	0.401	1.01
Weight on Payments Baseline α_B	0.902	0.906	0.042	0.819	0.979	1.01
Weight on Payments Treatment α_T	0.863	0.867	0.056	0.753	0.966	1.01
Weight on Preferences Baseline γ_B	0.837	0.848	0.084	0.671	0.982	1.01
Weight on Preferences Treatment γ_T	0.779	0.788	0.106	0.575	0.971	1.01
Prior Treshold Baseline δ_B	0.958	0.96	0.08	0.79	1.112	1.01
Prior Treshold Treatment δ_T	0.946	0.945	0.107	0.73	1.152	1

(a) Posterior parameter summary



(b) Average and predicted choices

Figure 4: Summary Probabilistic Model Altruistic Choices. **(a)** Estimated parameter values of equation 7 based on 10000 posterior samples (+ 2000 warmup) per each of four chains. Parameters correspond to mean of log-normal hyper-distributions. Mean, median and sd refer to the mean, median and standard deviation of the posterior distribution samples. HDI 2.5% and HDI 97.5% indicate the borders of the 95% highest-density interval (HDI). $\hat{R}$ is a diagnostic of convergence of the Markov chains Markov chains ($\hat{R} = 1$ indicating convergence). **(b)** Average and predicted choices, including 95% HDI intervals.

The Table in panel (a) of Figure 4 summarizes the parameters of the model. Given the hierarchical nature of the model, one can inspect parameters both on the population level, i.e., the mean of the log-normal hyper-distribution of a given parameter, as well as individual parameters for each subject. Here, I focus on the former, Table A.8 in the appendix shows the resulting (average) individual parameters.[18] Both tables contains the median and mean of the posterior samples of the respective parameter, the 95 % credible interval, the shortest interval containing 95 % of probability mass as well as

[18]The accompanying online appendix plots the individual choice curves and the individual data for each subject: https://nmwitzig.github.io/noise-app.html

the $\hat{R}$ convergence diagnostic (Vehtari et al., 2021) with $\hat{R} < 1.05$ often considered as necessary condition.

I first focus on the "base parameters", i.e., parameters that do not differ by treatment group. First, the altruistic preference parameter $\beta = 0.338$ $[0.304 - 0.373]$ aligns with the behavior described previously: on average, participants approximately weigh their payment twice as important compared to the payment of the other person. The probabilistic model also yields estimates of the *prior parameters*: The outcomes prior representing the intuitive perception of monetary payments is parameterized with the following mean $\mu^{\hat{r}} = 0.964$ $[0.503 - 1.492]$ and standard deviation $\sigma^{\hat{r}} = 0.930$ $[0.495 - 1.536]$, and the preference prior representing intuitive preferences by $\mu^{\hat{b}} = 0.816$ $[0.299 - 1.560]$ and standard deviation $\sigma^{\hat{b}} = 0.716$ $[0.079 - 2.097]$. Using the posterior samples, I can inspect the numerical relationship between these parameters: Constructing auxiliary variables $\sigma^{\hat{r}} - \sigma^{\hat{b}}$ and $\mu^{\hat{r}} > \mu^{\hat{b}}$ yields the moderate probabilities that $P(\sigma^{\hat{r}} > \sigma^{\hat{b}}) = 0.732$ (which would translate into a stronger reliance on the preference prior), and $P(\mu^{\hat{r}} > \mu^{\hat{b}}) = 0.736$, indicating a higher level of "selfishness" in the preference prior compared to the "pro-sociality" of the preference prior.

Recall that the primary goal of the experiment was to increase noise levels in the Treatment group. This has been successfully achieved according to the probabilistic model: Noise is smaller in the Baseline $\nu_B = 0.277$ $[0.232 - 0.326]$ compared to the Treatment group $\nu_T = 0.338$ $[0.281 - 0.401]$ with $P(\nu_B < \nu_T) = 0.999$. Further, the hyper-parameters of the noise distribution (on the log scale) differ between condition as $\mu^{\nu}_T - \mu^{\nu}_B = 0.198$ $[0.074 - 0.332]$ (see Table A.7). This translates into smaller values of α and γ in the Treatment group as $\alpha_B = 0.902$ $[0.819 - 0.979]$, $\alpha_T = 0.863[0.753 - 0.966]$,$\gamma_B = 0.837[0.671 - 0.982]$ and $\gamma_T = 0.779$ $[0.575 - 0.971]$. As discussed in Section 2, lower values of γ (relative to α) can lead to a "mechanical" shift towards more pro-social choices, as any given preference threshold will be "perceived" larger under noise. This is a therefore a first candidate explanation for the origin of the treatment effect. However, given that $P(\alpha_T - \gamma_T > \alpha_B - \gamma_B) = 0.731$, the model provides only moderate evidence for such a mechanical increase in the Treatment group.

In addition to prior variances, (differences in) prior means can shift behavior, too: With $\mu^{\hat{r}} > \mu^{\hat{b}}$, the prior threshold δ becomes – inevitably – smaller than 1 (i.e., a negative value on the log scale) in both groups: $\delta_B = 0.958$ $[0.79 - 1.112]$ and $\delta_T = 0.946$ $[0.73 - 1.152]$ with $P(\delta_B < 1 = 0.761)$ and $P(\delta_T < 1 = 0.760)$. This, in turn, implies a bias towards *more* selfish choices. This is in contrast to the mechanism described above and to the observed treatment effect.

While a stronger reliance on the preference prior thus remains a potential candidate for the mechanism of the treatment effect, the fact that the (noise-induced) threshold points towards *more* selfish choices, renders the results of the probabilistic model ambiguous and it thus remains less clear where the differences in Baseline and Treatment Group originate from. Furthermore, what the discussion so far largely neglected is the considerable degree of uncertainty surrounding the prior moments (see sd column), which is also reflected in the largely moderate differences in prior moments (and derived quantities and probabilistic statements). This cautions against interpreting too much based on the mean parameter values. Further reason for caution is provided in Figure A.7. There, a model comparison compares the predictive power of the "full" model (equation 7) with simpler variants of it.[19] While the "full" model provides the highest $ELPD_{WAIC}$ among

[19] More specifically, I formulate models that either allow only for (i) differences in prior means fixing

all models considered, the difference in goodness-of-fit compared to the simpler models are relatively minuscule. Although all variants of the "full model" outperform the standard random-utility benchmark – supporting the overall modeling approach – the small differences in predictive power between the variants reinforce the earlier conclusion that the altruism data is insufficient to conclusively identify the exact origin of the treatment effect.

Result 2: *The exact origin of the treatment effect in the altruistic choices, i.e., its mechanism, cannot be conclusively identified based on the altruism data alone.*

4.3 Number Comparison Task: Theory and Descriptives

As stated, the evidence thus far does not conclusively inform about the origins of the treatment effect in the altruistic choices. Therefore, I further investigate the origins of the treatment variation with the data from the number comparison task. This data is insightful as the choice of the number comparison shares similar "mechanics" with the altruistic choices (comparing two numbers), while abstracting from any subjective altruistic preference. Akin to equation 7, the choices in the number comparison task can be understood as a result of the following choice function:

$$Pr[(A \succ B \times 1/2] = \Phi\left(\frac{\alpha^{'} \times \ln\left(\frac{A}{B}\right) - \ln(\frac{1}{2}) - \ln(\delta^{'})}{\nu^{'}\alpha^{'}}\right) \tag{9}$$

where $\frac{A}{B}$ is the ratio of numbers A and B (previously *self* and *other*), $\alpha^{'} = \frac{\sigma^{\hat{r}'2}}{\sigma^{\hat{r}'2}+\nu'^2}$ and $\delta' = \frac{1}{\mu^{\hat{r}'1-\alpha'}}$. Equation 9 assumes that the term $1/2$ – an objectively stated constant – is perceived without noise (as opposed to the term $\frac{\beta}{1-\beta}$ in equation 7). Assuming that the treatment variation works similarly across domains (and that this functional form is appropriate), investigating the number comparison data allows to compare between H_{1a}, H_{1b} and H_{1c}: if H_{1a} is the driver behind the treatment effect, i.e., a stronger reliance on a pro-social prior (and maintaining the assumption that $\mu^{\hat{r}} = \mu^{\hat{b}} = 1$), higher levels of ν will *increase* average choices for A in the number comparison: Absent the influence of their altruistic preference, more noise will again mechanically bias $\frac{A}{B} < 1$ to larger values on the original scale. Leveraging the intuition employed above, this can be also understood as a bias toward an intuition that $\frac{\hat{A}}{\hat{B}} = 1$(see last row of Figure A.2). In contrast, under H_{1a}, i.e., if subjects intuitively adapt towards some intermediary $0 < \frac{\hat{A}}{\hat{B}} < 1$, more noise will *decrease* average choices for A, as larger values of $\frac{A}{B}$ will appear smaller (see first two rows of Figure A.2). Under H_{1c} equation 9 reduces to $P(A) = \Phi\left(\frac{\ln\frac{A}{B} - \ln\frac{1}{2}}{\nu \times \sqrt{2}}\right)$, which results in

a common prior variance η^2, (ii) differences in prior variances with a common prior mean κ, or (iii) assume identical prior means and variances which corresponds to a "pure noise" model. For reference, I also include a (iv) standard random utility benchmark (see Table A.11). I inspect $ELPD_{WAIC}$ values which measure the goodness of fit minus a model complexity penalty (Watanabe, 2013) and provide a computationally less demanding approximation to leave-one-out out-of-sample prediction accuracy (Vehtari et al., 2017), while accommodating model uncertainty. The worst variant of the simpler model is a "Pure Noise" model with a $\Delta ELPD_{WAIC} = 15.76$, although this difference – given the relatively large standard errors – does not appear too meaningful.

on average slightly fewer choices for A, while the indifference value between both groups remains unchanged.

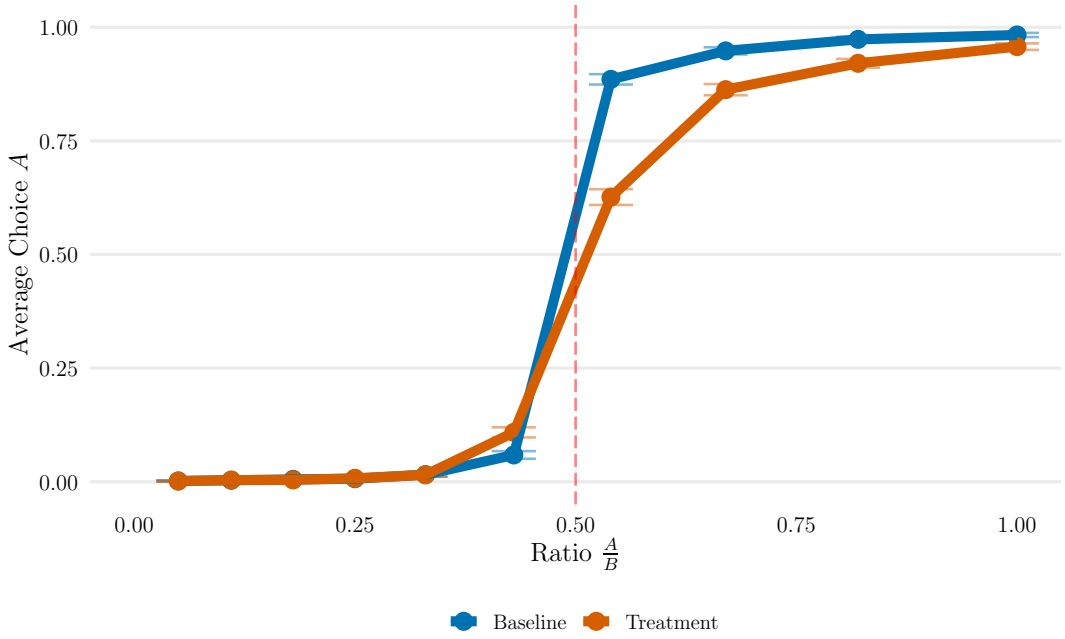


Figure 5: Choices for A in Baseline and Treatment group. This plot shows the association between average choice for A and distinct values of the ratio $\frac{A}{B}$, separately drawn for the Baseline and Treatment group with 95% confidence intervals around mean values.

The group differences in behavior in the number comparison task are shown in Figure 5, which plots the average choices for A as a function of $\frac{A}{B}$, separately drawn for Baseline and Treatment. To maximize payoffs, subjects should choose B whenever $\frac{A}{B} < 0.5$ and choose A whenever $\frac{A}{B} > 0.5$ (vertical red dashed line at $\frac{A}{B} = 0.5$). This data offers several insights: First, the Baseline group again shows a steeper association between choices and changes in the values of $\frac{A}{B}$ compared to the Treatment group. This translates into the Baseline group identifying the correct solution in 96,98% of trials compared to 92.26 % in the Treatment group ($p < 0.001$). The Treatment group also decides less often for A (i.e., thus errs asymmetrically): $\bar{A}_B = 0.388$, $\bar{A}_T = 0.351 (p < 0.001)$. Both observations are confirmed by a linear probability model in Table A.4, which tells that the Treatment group decides 3.7 percentage points less for A and an increase in $\frac{A}{B}$ by 1 has a 9 percentage points lower effect in the Treatment group compared to the Baseline (similar conclusions are drawn based on a Probit model in Table A.5). Another apparent observation is that choices are much more *consistent* in the number comparison task than altruistic choices. This is unsurprising given that I fix a common threshold term of $1/2$ replacing an individual-specific determined preference threshold, which eliminates choice differences due to individual heterogeneity in altruistic preferences and noise in its perception.[20] As $1/2$ is an *objective* quantity, I expect less subjective uncertainty around its value, which in turn also translates to higher choice consistency (note that equation 7 contains two sources of uncertainty whereas equation 9 contains only one).

[20] See Section 4.5 for a more detailed discussion on (separate) noise terms in altruism and number comparison.

Result 3: *In the number comparison task, The Treatment group again shows flatter association between changes in numerical magnitudes and choices and decides less often for A.*

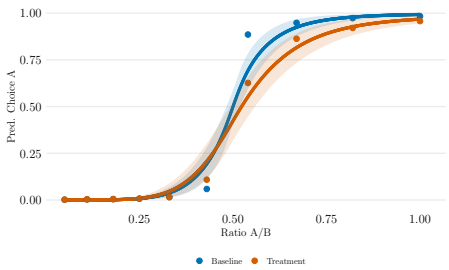
Transporting these findings to the previous results in altruism choices – and assuming the treatment variation works similarly across tasks – a common explanation for the treatment effect in both groups would be the mechanism underlying H_{1b}: Participants rely on an intuition that $0 < \frac{\widehat{A}}{\widehat{B}} < 1$ (and $0 < \frac{\widehat{self}}{\widehat{other}} < 1$), which in particular biases the perception of larger ratios downwards which then turns the Treatment group towards fewer choices for A and $self$.

4.4 Number Comparison Task: Probabilistic Model

I can also estimate a probabilistic model based on the number comparison data equivalently to Section 4.2 and investigate the probability of the parameter values of equation 9. The estimated parameters are shown in panel (a) of Figure 6 with average individual parameters in Table A.10.

	mean	median	sd	hdi 2.5%	hdi 97.5%	$\hat{R}$
Base Parameters:						
Prior Std. Outcomes $\sigma^{\hat{r}'}$	0.774	0.772	0.118	0.544	1.005	1.03
Prior Mean Outcomes $\mu^{\hat{r}'}$	0.483	0.466	0.11	0.393	0.603	1.01
Group Specific:						
Noise Baseline ν'_B	0.182	0.182	0.009	0.165	0.2	1
Noise Treatment ν'_T	0.258	0.258	0.013	0.233	0.283	1
Weight on Numbers Baseline α'_B	0.944	0.947	0.017	0.909	0.973	1.03
Weight on Numbers Treatment α'_T	0.894	0.899	0.032	0.83	0.948	1.03
Prior Treshold Baseline δ'_B	1.043	1.042	0.016	1.013	1.078	1.03
Prior Treshold Treatment δ'_T	1.084	1.081	0.032	1.023	1.15	1.02

(a) Parameters posterior summary



(b) Average and predicted choices

Figure 6: Model Summary Number Comparison **(a)** Estimated parameter values of equation 7 based on 10000 posterior samples (+ 2000 warmup) per each of four chains. Parameters correspond to mean of log-normal hyper-distributions. Mean, median and sd refer to the mean, median and standard deviation of the posterior distribution draws. HDI 2.5% and HDI 97.5% indicate the borders of the 95% highest-density interval (HDI). $\hat{R}$ is a diagnostic of convergence of the Markov chains Markov chains ($\hat{R} = 1$ indicating convergence). **(b)** Average and predicted choices, including 95% HDI intervals.

Given the assumed functional form, the base (treatment-group invariant) parameters now only consists of the moments of the numerical magnitude prior. This prior is parameterized by $\sigma^{\hat{r}'} = 0.774$ $[0.544-1.005]$ and $\mu^{\hat{r}'} = 0.483$ $[0.393-0.603]$, i..e, an "intermediate" intuitive perception of numerical magnitudes, that $\frac{\widehat{A}}{\widehat{B}} < 1$. Before interpreting the impact of this prior on choices in more detail, it is useful to inspect the differences in noise(-related) parameters across groups first: Again, noise is smaller in the Baseline versus the Treatment group as $\nu'_B = 0.182$ $[0.165 - 0.2]$, $\nu'_T = 0.258$ $[0.233 - 0.283]$ and $P(\nu'_T > \nu'_B = 1)$. Note also that $\mu^{\nu'}_T - \mu^{\nu'}_B = 0.348$ $[0.200 - 0.484]$ (c.f. Table A.9). Given larger noise levels in the Treatment group, the weight on payments is consequently smaller with $\alpha'_T = 0.894$ $[0.83 - 0.948]$, $\alpha'_B = 0.944$ $[0.909 - 0.973]$. Compared to the values of ν from Figure 4, ν' values are thus overall lower in the number comparison task while the *treatment effect* is more pronounced compared to the altruism domain.

The *origins* of this treatment effect, in contrast to previously, are now much more conclusively identified. Based on the posterior samples, I calculate that $P(\mu^{\hat{r}'} < 1 = 0.929)$

, lending strong support to an 'intermediate" intuitive numerical magnitude perception. Importantly, this translates into $P(\delta'_B > 1 = 0.929), P(\delta'_T > 1 = 0.929)$, too, which implies strong evidence for a (noise-induced) bias towards *fewer* choices for A in both groups (matching the direction of the treatment effect), which further is *larger* in the Treatment compared to the Baseline group given $P(\delta'_T > \delta'_B) = 0.996$.

Overall, the probabilistic model based on the number comparison data thus yields a clearer indication how the treatment effect operates, namely through a biased perception of numerical magnitudes. This leads to in particular larger values of $\frac{A}{B}$ being perceived as smaller under noise and thus a bias towards larger values more generally. This is in contrast to the much more moderately sized potential (and ambiguous) explanations for the treatment effect discussed in Figure 4 and the much larger HDI intervals surrounding the moments of the priors there.

Result 4: *The probabilistic model of the number comparison behavior indicates a high probability for a presence of an "intermediate" perception of numerical magnitudes as the driver of the treatment effect.*

The individual choice curves depicted in panel (b) of Figure 6 show that the average choices are by and large close to the HDI areas indicating that the structural estimates can reasonably recover the average behavior by participants. Figure A.6 further plots average and predicted choices for A for each individual with a rank-correlation of $\rho = 0.94$ between the two. However, especially in the Baseline group, the intervals sometimes do not include the average behavior, which indicates that the chosen functional form can not fully explain these data points. However, in comparison with a "pure noise" model (the model corresponding to the mechanism underlying H_{1c}, i.e., assuming identical prior variances for *both* the perception of numbers $\frac{A}{B}$ as well as the term $1/2$, the model specification in equation 9 is superior as indicated by the $ELPD_{WAIC}$ values (see Figure A.8). This strongly suggests that a single estimated parameter ν is unable to capture the behavior of our subjects in the number comparison task and suggests that an additional driver, here a numerical magnitude prior with an "intermediate" mean is at play.

If I transport the findings from the number comparison data to the altruistic choices, H_{1a} appears most likely to be responsible for the treatment effect, which would be interpreted as follows: Under the treatment variation, participants relied relatively more strongly on an intermediate intuitive perception that $self < other$, i.e., that the payment prior mean $0 < \mu^{\hat{r}} = \frac{\widehat{self}}{\widehat{other}} < 1$. This fit to the nature of the treatment variation: Encasing monetary payments in to-be-calculated sums instead of showing plain values predominantly biases the perception of monetary payments instead altruistic preferences.

4.5 Identification, Noise in Altruism and Nature of the Treatment Effect

A caveat to this interpretation remains, however, that it requires an explicit "logical transfer" from the number comparison to the altruism domain. This is related to the *identification* of the model, which is not fully given based on only the altruism data and thus may characterize a weakness of the model. While identifiability has a different connotation in Bayesian models compared to the a classical understanding, one intuitive way to think about identifiability is the the difference between the prior and posterior (parameter) distribution, i.e., how informative the data is (see e.g., Xie and Carlin 2006).

Prior predictive checks in A.5 show that the average parameter values and corresponding behavior based on the chosen priors is "sufficiently" different to conclude that the altruism data informs the posterior distributions, particularly for ν and β yet to a lesser extent for the prior parameters $\mu^{\hat{b}}$, $\sigma^{\hat{b}}$, $\mu^{\hat{r}}$, $\sigma^{\hat{r}}$.This calls into doubt whether (modelling) noise in altruism is then actually necessary to explain altruistic behavior in the present setting and if – instead – only noise in perceiving numerical magnitudes is sufficient to explain behavior. This seems especially pertinent given that the treatment effect likely operates through biased numerical magnitude perception as discussed above.

To investigate this critique, I take a more principled approach compared to a mere "logical transfer" and combine the datasets from both tasks to *jointly* estimate the parameters of equations 7 and 9. In the combined estimation, the parameters $\nu, \mu^{\hat{r}}, \sigma^{\hat{r}}$, are inferred from both number comparison *and* altruistic choices, whereas $\sigma^{\hat{b}}$, $\mu^{\hat{r}}$ and $\frac{\beta}{1-\beta}$ are estimated from only the altruistic choice data.For model estimation, this simply requires to include two likelihoods, again demonstrating the high flexibility of Bayesian models. The resulting parameter values of the combined estimation are shown in Table A.13. The combined model confirms the conclusions drawn previously: With $P(\mu^{\hat{r}} < 1 = 1)$, $P(\sigma^{\hat{r}} < \sigma^{\hat{b}} = 0.999)$, and, given $P(\mu^{\hat{r}} < \mu^{\hat{b}} = 0.710)$ and in turn, $P(\delta_T > 1 = 0.999)$, $P(\delta_B > 1 = 0.999)$, the combined model provides strong support in favor of a (noise-induced) biased "intermediate" numerical perception in favor of fewer choices for *self* and *A*. Compared to Figure 4, these stronger probabilistic statements are due to much tighter posterior distributions, especially around the prior moments, speaking in favor of an increase identifiability of the jointly estimated model.

In addition to a higher degree of identifiability, the combined dataset also allows to address the above-mentioned critique that perhaps only numerical magnitudes are perceived with noise, yet not altruistic preferences. This can be easily modeled accordingly with a simpler variant of the "full model". However, compared to the "full model", such a simpler model performs considerably worse in explaining altruistic choices (see Figure A.9), which implied that noise in number perception only is insufficient to explain altruistic choices of participants. This argument is further supported by a simple linear regression provided in Table A.6 which shows that, given participant and game fixed effects (which account for individual differences in altruism and noise), the inconsistency in a given trial is significantly higher in the altruism compared to the number comparison task.

Lastly, the combined dataset further allows to address a critique that calls into doubt whether noise is *domain-invariant*, i.e., doubts the appropriateness of the (simplifying) common noise assumption. In Section A.3.2, I describe and estimate a model that incorporates distinct signal variances $\nu^2_{\frac{self}{other}}$ and $\nu^2_{\frac{\beta}{1-\beta}}$. This model allows for several probabilistic statements of previously carried out interpretations: For example, recall that estimated noise was overall smaller in the number comparison data compared to the altruism data (cf. Figures 4 and 6). Based on the combined model, I can estimate that $P(\nu_{\frac{\beta}{1-\beta},B} > \nu_{\frac{self}{other},B}) = 1$ and $P(\nu_{\frac{\beta}{1-\beta},T} > \nu_{\frac{self}{other},T}) = 1$, which reverberates the previous finding. Furthermore, the *treatment effect* was larger in the number comparison compared to the altruism domain. Again, the model with separate noise terms *domain-specific* treatment effects and permits their probabilistic comparison: with $\mu_T^{\nu_{\frac{self}{other}}} - \mu_B^{\nu_{\frac{self}{other}}} = 0.327$ [0.210 − 0.446], $\mu_T^{\nu_{\frac{\beta}{1-\beta}}} - \mu_B^{\nu_{\frac{\beta}{1-\beta}}} = 0.187$ [-0.105 − 0.478] and $P(\mu_T^{\nu_{\frac{self}{other}}} - \mu_B^{\nu_{\frac{self}{other}}}) > (\mu_T^{\nu_{\frac{\beta}{1-\beta}}} - \mu_B^{\nu_{\frac{\beta}{1-\beta}}}) = 0.804$, the model again strengthens the previous interpretation. Note further that given $P(\mu_T^{\nu_{\frac{self}{other}}} - \mu_B^{\nu_{\frac{self}{other}}} > 0 = 0.892)$, this model

demonstrates how encasing the values of *self* and *other* in to-be-calculated sums can have an impact on perception of altruistic preferences, too which yet remains smaller compared to the impact on perceiving numerical magnitudes. One caveat to this model remains in that estimating separate noise terms comes at the cost of identifiability and precision (see larger values of the posterior standard deviation in Table A.14 compared to Table A.13).

4.6 Alternative Explanations for the Treatment Effect

On a more critical note, one might argue that some other (unintended) effect of the chosen variation is responsible for the differences between treatment groups beyond the mechanisms proposed by the theoretical model.

In Section A.3.3, I discuss several alternative explanations. For example, one argument could be that the treatment effect (i.e., the aforementioned intuition) is "learned" over the repeated trials of the experiment (which in turn could limit the external validity of our results). However, treatment differences towards more pro-sociality already materialize in the initial 10 (hypothetical) practice trials, I also do not find evidence for a (growing) difference in fatigue as the driver of the treatment effect and provide arguments against a purely "mechanical" increase in pro-sociality due to how the sums of the Treatment group were constructed. Furthermore, in Section A.3.4, I estimate heterogeneous treatment effects and shows that the treatment variation did not work systematically differently for participants, who e.g., expect or hold different norms between the Baseline and Treatment variant of our task. Most personal characteristics do not meaningfully contribute to heterogeneity in the treatment effect, if anything, the treatment effect is slightly weaker for participants with high levels of self-reported altruism and "Need for Cognition".

Overall, the best guess on the origin of the treatment effect is therefore that participants quickly understand how the task works: "less-for-me" vs. "more-for-other", which is reflected in their intuitive perception of the respective monetary payments and magnitudes. I will return to this point and its potential implications in more detail in Section 5.

4.7 Altruism, Number Comparison and Cognitive Ability

I now explore H_2 and H_3, i.e., whether behavior in altruistic choices and number comparison is correlated and if measures of cognitive ability correlate with individual measures of cognitive noise ν_i.

4.7.1 Altruistic Choices and Number Comparison

As stated earlier, if similar cognitive processes (partly) guide altruism and number comparison choices, I expect to see some association between behavior in both tasks. Given that the numbers featured in the trials of the altruism choices and the number comparison task are *identical*, I can closely examine such possible relationships. Using a linear probability model, Table 1 explores if choices for *self* in the altruism choices are correlated with future choices in the number comparison task *in the exact same trial*.[21] With multiple iterations per group of trials, I can further include both participant- and game-

[21] As I have five repetitions of each unique trials in both tasks, I additionally need to match the data between tasks on an occurrence variable, which tracks in which order a participant encountered a given trial in a given game (i.e., a group of identical trials). I thus match the choices of the first altruism trial of a given game to the first number comparison choice in the same game, and so on.

(one game consisting of the 5 repetitions of a given trial) fixed effects, with the former including the treatment effect.

	(1)	(2)	(3)
A chosen	0.048***		0.045***
	(0.008)		(0.010)
Correct Number Comparison		0.029***	0.006
		(0.009)	(0.010)
Num.Obs.	60000	60000	60000
Participant Fixed Effects	Yes	Yes	Yes
Game Fixed Effects	Yes	Yes	Yes
Clustered Standard Erorrs	Yes	Yes	Yes
Unique Obs	300	300	300
R2	0.001	0.000	0.001

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

Table 1: Correlation of Choices between Tasks. Linear Probability Model. Dependent variable is the choice for $self$. Clustered standard errors (participant-level) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Column (1) shows that if a person will choose A in a given trial, they are 4.8 percentage points more likely to choose $self$. Moreover, by Column (2), more *correct* choices in a given trial also positively correlate with choices for $self$: A person is 2.9 percentage points more likely to choose $self$ if they will identify the correct solution in the number comparison in that trial. This implies that factual errors in the number comparison correspond to more altruistic choices, which in turn fits to the Treatment group being both more altruistic and making more errors in the number comparison task (recall that I control for treatment differences with participant fixed effects). However, the correlation between correct and selfish choices vanishes once I include the choice for A in column (3). This can be explained by the fact that both the Baseline and Treatment group errs on the side of A, i.e., chooses A not often enough (see Figure 5).

Inconsistencies across tasks are (moderately) correlated, too: The average standard deviations in the altruism task and the number comparison are positively correlated in both the Baseline ($\rho = 0.256$, $p = 0.0015$) and the Treatment ($\rho = 0.139$, $p = 0.089$) group. Participants who are more inconsistent in their altruistic choices are thus also slightly more inconsistent in the number comparison task (yet to a smaller extent in the Treatment group). In addition, in the Treatment group, the individual noise parameter values across tasks, ν and ν' are positively correlated, but not in the Baseline group (Treatment: $\rho = 0.229$, $p = 0.004$; Baseline: $\rho = 0.103$, $p = 0.201$). Notably, the Treatment correlation coefficient of 0.229 is similar in magnitude to the reported rank correlation coefficient of 0.26 in Frydman and Jin (2022). There, a parameter n (that indicates the precision of the mental representation of monetary payoffs in their model) correlates between a risky lottery and a "perceptual" choice task in which participants had to identify if a given number shown is larger or smaller compared to some reference number.

The fact that the correlation is smaller in the Baseline group could be related to the fact that the number comparison task is relatively easy given sufficient time which in turn leads to a high choice consistency that somewhat mutes the impact of individual noise.[22]

[22]Note also that, in the number comparison task, participants know there exists a *correct solution* and

As Section 4.8.1 shows, thinking times – a common measure of decision difficulty – are positively correlated across tasks in *both* the Baseline and Treatment group.

Leveraging the multivariate normal setup of the separate noise terms model described in Section 4.5, I can further inspect correlations between (jointly) estimated noise terms across domains, which results in $\rho_{\nu \frac{self}{other} B, \nu \frac{\beta}{1-\beta} B} = 0.065$[-0.194−0.290] and $P(\rho_{\nu \frac{self}{other} B, \nu \frac{\beta}{1-\beta} B} > 0 = 0.703)$. In the separate noise model, (base) individual noise terms are thus moderately positively correlated.

Overall, I thus interpret the presented evidence as tentative support for H_2 and an interpretation that processes which guide imprecisions in number comparison also partly guide imprecisions in altruism choices if gathered in a similar way, although the link is not as straightforward as in previous work.

Result 5: *Behavior and choice inconsistency, as well as individual noise measures are moderately positively correlated in the altruism and number comparison task.*

4.7.2 (Self-reported) Cognitive Ability and Individual Measures of Noise

To investigate H_3, i.e., to test for a negative relationship between individual measures of cognitive ability and cognitive noise, I compute correlations between the CRT, BNT, and NFC scale as well as the preferences for intuition and deliberation scale with individual structural measures of noise both from altruistic choices (ν_i) and the number comparison (ν_i'). For further reference, I also include the altruistic preference parameter β_i as well as more altruism-related measures. Table 2 contains the rank correlation coefficients between the various measures and structural parameters:

	Altruism Noise ν	Altr. Preference β
Cognition-related:		
No. Correct CRT	-0.296***	0.180**
Berlin Numeracy Test	-0.273***	0.188**
Avg. Need for Cognition	-0.157**	0.145*
Avg. Deliberation	0.140*	-0.145*
Avg. Intuition	0.107	-0.103
Altruism-related:		
Dictator Game Other	-0.317***	0.475***
GPS Value Gift	-0.090	0.132*
GPS Donation	-0.029	0.128*

Table 2: Correlation Structural Parameters and Individual Characteristics. Need for Cognition, Deliberation and Intuition averaged values. p-values from pairwise rank-correlation tests ($n = 300$). * p < 0.1, ** p < 0.05, *** p < 0.01.

Focusing on the first column, I observe a negative correlation between the number of correct items in the Cognitive Reflection Task and individual measures of ν_i. Participants

even though taking longer reduces the eventual payoff, they often invest ample time to find the correct answer. This is a marked difference compared to the altruism choices, where no objectively correct solution exists.

who score better on the CRT exhibit a lower level of ν_i. Similarly, the higher the score on the Berlin Numeracy Test and the higher self-reported "Need for Cognition", the lower the individual estimate of noise ν_i. These associations – though only correlational – underscore an important point: The proposed theoretical model, and the ν parameter in particular, indeed appear to relate to a *cognitive* component of the process of making altruistic choices, which provides validating evidence for the overall approach. This is further in line with the above-mentioned work showing how CRT performance correlates with biases and mistakes in choices (Augenblick et al., 2022; Oprea, 2024; Assenza et al., 2019; Chew et al., 2022). In contrast, self-reported preferences for deliberation and intuition do not meaningfully correlate with ν_i.

Result 6: *Individual measures of cognitive noise negatively correlate with cognitive ability as measured by performance in the Cognitive Reflection Task and Berlin Numeracy Test.*

Table 2 further contains correlations with the values of the altruistic preference parameter β_i. These values correlate positively with the amount a participant gave to another person in the simple dictator game and also, albeit to a lesser extent, with the hypothetical gift-giving and donation decision from the GPS. Furthermore, β_i positively correlates with the CRT and BNT performance[23], whereas ν_i negatively correlates with the amount given in the simple dictator game. While I abstain from hypothesizing on the origins of this nexus, it could be related to the particular structure of the hierarchical model: The hyper-parameters of noise and altruistic preference distribution *negatively* correlate with one another as $\rho_{\nu B,\beta}$ = -0.676[-0.838 − -0.508], which in turn could explain the above-mentioned effects.

4.8 Response Times and Metacognition

I now turn to additionally study two core components insightful for choice processes: Response times and measures of "metacognition", which as understood here comprises several measures about participants' subjective thinking about their choices.

4.8.1 Response Time

I begin by investigating response times (RT), i.e., the amount of time a participant took to decide in both tasks. In psychology and cognitive science, RT is a highly-informative variable of the choice process (see e.g., Luce 1991) which established the following "standard results": for discriminating between stimuli, RT is higher the more similar the stimuli which is often attributed to a higher trial difficulty. This is true both for physical stimuli, such as the brightness of two lights (see e.g., Pins and Bonnet 1996, "Pierons Law"), as well as for numerical stimuli, such as two Arabic numerals (see e.g., Moyer and Landauer 1967.

For economic research, arguably the most important insight from RT stems from its close relationship to the strength-of-preference. Similar to the perceptual difficulty described above, the closer a subject is to indifference in an economic choice task, the longer their RT (see e.g., Alós-Ferrer and Garagnani 2022). RT has also been used to investigate

[23]But note that there is no correlation between CRT performance and the average choice for $self$: $\rho = -0.089, p = 0.123$.

social preferences, especially under the umbrella of dual-process model with fast (slow) decisions usually attributed to intuitive (deliberate) reasoning. Time pressure studies concluded that people intuitively tend towards cooperation (Rand et al., 2012; Rand and Kraft-Todd, 2014) ("Social Heuristics Hypothesis"), and that "fairness is intuitive" (Cappelen et al., 2016).[24] Recently, Bavard et al. (2024) show how humans are able to infer other people's social preferences only based on observing their RT.

In the present setting, I analyze RT and its correlation with behavior from various angles. First, a straightforward test is to investigate if our treatment variation, aimed at increasing the cognitive difficulty, actually lead to higher RT in the Treatment group. This is the case: On average, participants in the Baseline group took 1.33 seconds to decide in the altruism task, whereas participants in the Treatment condition took 2.02 seconds ($p < 0.001$). The difference is more pronounced in the number comparison task, with a mean RT of 1.36 seconds in the Baseline and 2.6 seconds in the Treatment group ($p < 0.001$). This supports the treatment intention that the sums increased the difficulty of deciding in both tasks. Within individuals, RT between the two tasks is (moderately positively) correlated both in the Baseline ($\rho = 0.25, p < 0.01,$) and Treatment ($\rho = 0.238, p < 0.01$) group. Participants, for which the altruism task was more difficult thus also had a higher difficulty of identifying the solution in the number comparison task (equating longer RT with choice difficulty).

I also investigate the distribution of RT and its relationship with the strength-of-preference. For this, the individual structural estimates outlined in Section 4.2 and 4.3 can be utilized: These estimates allow to infer the (mean) indifference values, i.e, at which value a subject is indifferent between $self$ and $other$ ($\widetilde{\frac{self}{other}}_i$), resp. A and B ($\widetilde{\frac{A}{B}}_i$). I can then calculate the difference of the ratio of a current trial j to that indifference value ($\Delta\widetilde{\frac{self}{other}}_{ij} = \frac{self}{other}_j - \widetilde{\frac{self}{other}}_i$) and investigate how their RT relates to that difference.

[24] This conclusion, however, has been challenged by subsequent work: Krajbich et al. (2015) show how such claims are often unwarranted once discriminability of choice options is accounted for. Similar findings are obtained by Merkel and Lohse (2019).

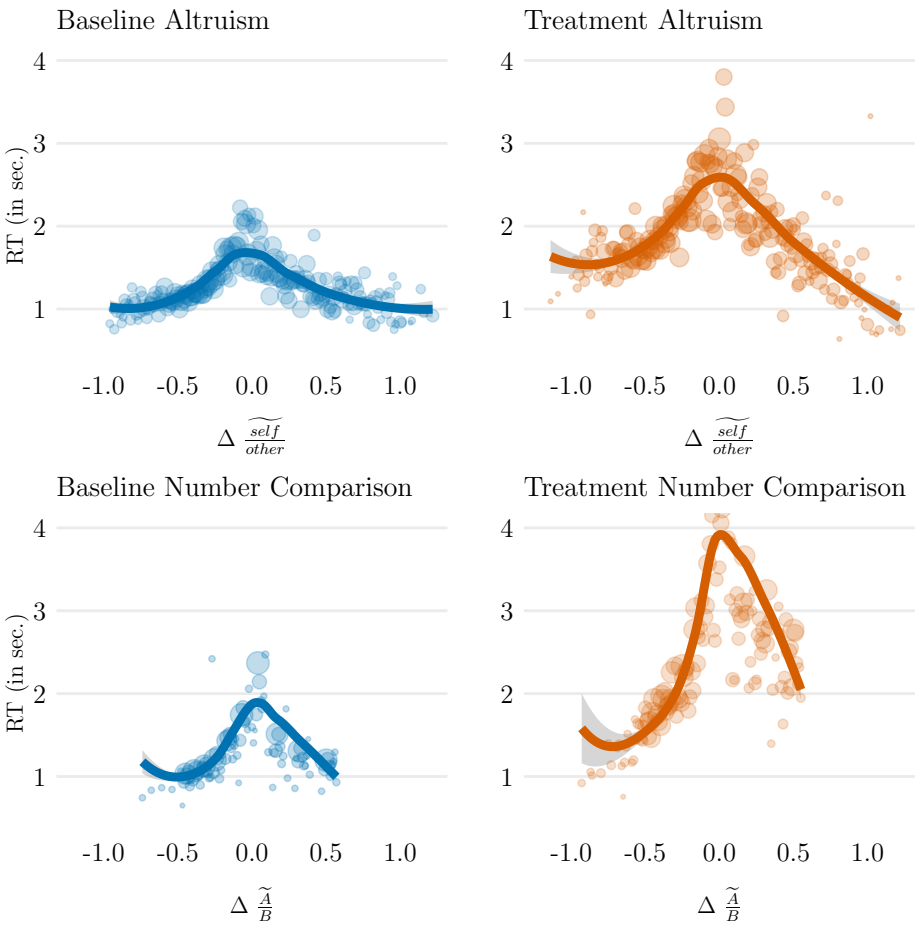


Figure 7: Distribution of RT and Distance to Predicted Indifference Ratio. $\Delta\widetilde{\frac{self}{other}}_{ij} = \frac{self}{other}_j - \widetilde{\frac{self}{other}}_i$, where $\widetilde{\frac{self}{other}}_i = \widetilde{\frac{self}{other}}_i : Pr(self_i, \widetilde{\frac{self}{other}}_i) = 0.5$ and $\frac{\widetilde{A}}{B}$ and $\Delta\frac{\widetilde{A}}{B}$ are constructed accordingly. The fit is from a local polynomial regression (with 95% confidence intervals). In addition, average data points are depicted with the size of the point proportional to its relative frequency.

Figure 7 plots the average RT (in a given trial) as a function of the difference to the individually predicted indifference value. For the data points, I aggregate over individuals according to the value of $\Delta\widetilde{\frac{self}{other}}$ and $\Delta\frac{\widetilde{A}}{B}$ and scale the size of the data points proportional to their relative frequency. The polynomials are fitted to the non-aggregated data. From this Figure, it becomes apparent that RT follows the usual pattern with its peak around the indifference value, i.e., that RT is largest at those ratios where the model predicts indifference.[25] I interpret this as validating evidence that the structural model and the proposed decision rule in Section 2 are useful in conceptualizing how subjects made their choices in both tasks and in understanding the respective decision difficulty.

Result 7: *Response Times are larger in the Treatment Group and largest where the probabilistic model predicts indifference.*

[25] Vieider (2024b) goes a step further and shows that the distribution of individual predictions of indifference exhibits a more pronounced pattern with RT compared to the expected value of lotteries, but in our setting, such a direct benchmark is not available (at least not for the altruism data).

RT and Choices I also investigate correlations of RT with choices. Table 3 contains 4 Probit models that regress choices for $self$ respectively A on the amount of RT. I log-transform the RT variable to reduce the impact of outliers (see e.g., Alós-Ferrer et al. 2016). I add participant fixed effects, which contain the treatment effect (columns 1 and 3) as well as game fixed effects (columns 2 and 4). In the first two columns, I observe a small positive and insignificant coefficient of the RT variable on the probability of choosing $self$. I thus do not observe a strong correlation between the amount of time a person took to decide and the level of altruism (even though the general treatment effect would also be consistent with a "fairness is intuitive" narrative). In contrast, I observe a pronounced positive relationship between RT and choices for A in columns (3) and (4). Longer RT is thus associated with a higher probability of choosing A, which could be interpreted that fast, intuitive answers lead participants to choose B, while more careful deliberation leads to A more often. In the altruism data, however, an analogous conclusion that quicker decisions are more often *other* (which would be in line with the interpretation of the treatment effect) is not warranted.

	(1)	(2)	(3)	(4)
$\frac{self}{other}, \frac{A}{B}$	5.160***		7.824***	
	(0.086)		(0.107)	
RT (log.)	0.034	0.020	0.482***	0.320***
	(0.021)	(0.022)	(0.018)	(0.019)
Data	Altruism	Altruism	Number Comp.	Number Comp.
Participant FE	Yes	Yes	Yes	Yes
Game FE	No	Yes	No	Yes
Num.Obs.	72000	72000	60000	60000
Clustered Standard Errors	Yes	Yes	Yes	Yes
Unique Obs	300	300	300	300

Table 3: Correlation of RT with Behavior. Probit Model. Columns (1) and (2) use data from the altruism choices and the dependent variable is the choice for $self$, columns (3) and (4) from the number comparison with choice for A as dependent variable. Clustered standard errors (participant-level, "bias-reduced linearization" Pustejovsky and Tipton 2018) in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Result 8: *Response Times correlate with behavior in the number comparison task, indicating that higher RT (i.e., more "deliberate" choices) corresponds to more choices for A.*

4.8.2 "Metacognition"

In addition to RT, I investigate the relationship between behavior and measures of "metacognition". i.e., measures on how participants think about their decisions. Recent literature shows how metacognition can play an important role in explaining (biases in) economic choices. Enke and Graeber (2023) show how self-reported cognitive uncertainty, i.e., "people's subjective uncertainty over which decision maximizes their expected utility" is predictive of a compression effect in various domains from risky choice to belief updating. Olschewski and Scheibehenne (2024) illustrate how information on metacognitive awareness of one's cognitive imprecisions improves Bayesian decision models in sample estimation tasks. Further, Oprea (2024) documents that self-reported measures of attention and noise correlate with prospect-theoretic behavior.

I can, too explore the links between (noise in) altruistic choices and number comparison and metacognitive self-reports. These comprise of a negative confidence (similar to cognitive uncertainty), attention, and the precision of comparison as well as additional belief-based measures from the number comparison task. I (i) first test for treatment differences in these metacognitive measures, and (ii) investigate their correlation (on a subject level) with the main choice data.

	Baseline (avg.)	Treatment (avg.)	p
Altruism:			
Negative Confidence	0.304	0.318	0.643
Avg. Attention	0.731	0.698	0.127
Precision	0.364	0.382	0.573
Number Comparison:			
|Δ Belief Correct|	0.084	0.163	<0.001***
Belief Correct Confidence	0.785	0.651	<0.001***
|Δ Belief Time Spent|	0.674	0.976	0.01***
Belief Time Spent Confidence	0.638	0.593	0.093*
Precision	0.568	0.447	<0.001***
Avg. Attention	0.761	0.705	0.007***

Table 4: Treatment Effects Metacognition. Note that 7 participants are omitted, where |Δ Belief Time Spent| > 10 in all tests (i.e., $n = 293$). p-values from two-sided t-test. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 4 displays the average values of various measures of metacognition, both from the domain of altruistic choices as well as number comparison, separately for Baseline and Treatment group alongside the p-value of a two-sided t-test. The first set of measures in the table contains self-reported confidence (how confident subjects are they made the for-them correct decision), the average attention (a subject paid to the values of $self$ and $other$), and precision, (i.e., if participants compared the payments more approximatively or did a precise comparison). The (negative of the) confidence measure is very similar between both groups with an average of 0.304 in the Baseline and 0.318 in the Treatment ($p = 0.643$).[26] For both the self-reported average attention (Baseline: 0.731, Treatment 0.698, $p = 0.127$) and precision (Baseline: 0.364, Treatment 0.382, $p = 0.573$) there is also no group difference.

This is markedly different in the number comparison domain: Here, participants in the Treatment group deviate more strongly in their beliefs from their true performance[27], i.e., have a larger |Δ Belief Correct| (Baseline: 0.084, Treatment 0.163, $p < 0.01$), report lower confidence in these belief statements (Baseline: 0.785, Treatment 0.651, $p < 0.01$), deviated more strongly in their belief how much time they think they needed in the number comparison (Baseline: 0.674 sec., Treatment 0.976 sec., $p = 0.01$), again report lower confidence in these estimates (Baseline: 0.638, Treatment 0.593, $p = 0.093$), also report a lower precision (Baseline: 0.568, Treatment 0.447, $p < 0.01$) and finally, lower

[26]This, in turn, is similar in magnitude to average cognitive uncertainty measures from typical lottery or balls-and-urns tasks (see Enke and Graeber 2023 and Amelio 2022).

[27]See also Figure A.12, which plots participants' beliefs of the average of correct answers (time spent) and their actual share of correct answers (time spent). Participants consistently underestimate the amount of correctly solved tasks *and* overestimate the amount of time spent, which results in a strong pessimistic bias in the beliefs of the number comparison task.

average attention (Baseline: 0.761, Treatment 0.705, $p < 0.01$). Similarly, correlations between choices, RT as well as inconsistencies are overall more pronounced in the number comparison domain compared to the altruism choices (see Table A.19 in the appendix).

Result 9: *Measures of "Metacognition" exhibit a strong treatment effect only in the number comparison task and not in the altruism task. Further, correlation between behavior and metacognition is more pronounced for number comparison than for altruism.*

5 Discussion of Results and Next Steps

I established the following main results: (i) encasing monetary payments in to-be-calculated sums caused more pro-social choices in the altruism task. This effect most likely operates through the perception of the payment values, as the effect manifests comparably in the number comparison task. (ii) I observe correlations in behavior between altruism choices and number comparison, (iii) a positive association between individual measures of cognitive noise and cognitive ability, and finally, (iv) a link between RT, metacognition and choices, which both reacts more strongly to the treatment and is more more pronounced in the number comparison compared to the altruism domain. I discuss each result in more detail, outlining potential avenues for future research in turn.

(i) Implications of the Treatment Effect A similar treatment effect in the altruism and number comparison task demonstrated how an intermediate intuition of $\widehat{A} < \widehat{B}$ (and $\widehat{self} < \widehat{other}$) is a candidate driver of the group differences. The exact origins of this intuition are less clear, but a possible explanation could be an instinctive understanding of the "rules" of the task, i.e., "less-for-me" vs. "more-for-other". Not only does our task carry such simple (and easy to grasp) "rules", it consequently fits also to the statistical environment of the tasks of the experiment. The empirical average ratio in the average trial amounts to $\frac{\bar{self}}{other} = 0.466$ and $\frac{\bar{A}}{B} = 0.437$. While an *adaption* to the statistics of both task thus remains a possible explanation for the origin of the treatment effect, the fact that treatment differences are already present during the very first trials and that I do not find clear evidence of learning effects (cf. Section 4.6), a very quick intuitive understanding of the task seems more plausible.

Either way, this challenges a common assumption in the noisy cognition literature that perceptions of monetary payments of different choice options are intuitively perceived to be the same (see e.g., Khaw et al. 2021). As in our tasks, other tasks could inherently imply certain statistical proportions and also allow for an instinctive understanding of the numerical relationship of stakes, e.g., in intertemporal decision-making ("smaller-and-sooner" vs. "larger-and-later"), or in lotteries, where risky and safe payoffs are necessarily different.[28]. This related to a point in Oprea and Vieider (2024, 33) who explicitly discuss differences between "naive" versus more sophisticated decision makers when specifying the prior mean parameter. Here, we provide strong support for the presence of such non-naive intuitions. This has important implications: If people quickly grasp "rules" or statistics of such tasks, this can potentially influences alter the direction of an increase

[28]However, note that e.g., in Vieider (2024b) the objects of perception are "benefits" and "costs" of risky and safe payoffs, where an intuitive understanding of them being equal is perhaps more convincing.

in cognitive noise.[29] While Khaw et al. (2021) attribute risk-aversion to higher levels of cognitive noise (and Barretto-García et al. (2023) show the neurological underpinnings of the model), a *causal* test of this direction is still to be performed to identify how behavior actually reacts to an increase in noise. Our treatment variation is a candidate for such a causal test.

However, I fully acknowledge that the interpretation of an intuition reminiscent of the "rules" of our task is so far purely speculative and not the result of an empirical test. Future work could therefore investigate the drivers of such a potential adaptation and e.g., exogenously manipulate choice environments that induce differences in intuitions (akin to efficient coding studies such as Frydman and Jin (2022); Polanía et al. (2019)) or explicitly model the noisy learning process (see Poggi (2021) for a start).

In addition, while I demonstrate *theoretically* how subjective uncertainty in one's subjective altruistic preference can affect choices, the implemented to-be-calculated sums likely operate primarily through numerical magnitude perception. While a model featuring separate noise terms showed that the chosen treatment variation may have also affected preference "perception", its impact was definitely smaller compared to its impact on numerical magnitude perception. An important next step would thus be to either design and implement an variation that exclusively affects the uncertainty in the perception of one's altruistic preference (without affecting numerical perception) or to compare the present treatment results with more standard time pressure or cognitive load treatments.

(ii) Correlation of Altruism and Number Comparison Our second set of results shows a (moderate) correlation between behavior and measures of noise in altruism choices and number comparison: I both observed an increase in choices for $self$ if a person chooses A in a future "twin"-trial and also if this person will identify the correct solution later on. Participants who are more inconsistent in choosing between $self$ and $other$ are also more inconsistent in the number comparison task. Both facts point towards some common driver between both domains, for which the noisy representation of monetary payments is a potential candidate. This is similar to conclusions in Frydman and Jin (2022) and Barretto-García et al. (2023), although the relationship between economic choice and numerical perception is weaker in our setting.[30] Nonetheless, common to both domains is the ability to *compare*, which in turn could be related to common cognitive processes. Note hat I specifically designed both tasks to be similar to each other. In turn, if such relationships between economic choice and numerical perception across domains manifest in other settings or what characteristics of a chosen setting determine the relationship remains a largely open question and seems worthy of future investigation.[31]

(iii) Cognition and Altruistic Preferences Our third result shows a correlation between measures of cognitive noise and cognitive ability. I interpret this as follows: Cognitive processes are likely to play a role at *expressing* one's (subjective) preferences, yet

[29] But also note that an "overfitting" of prior intuitions to a given statistical environment is not necessarily a given and not a good strategy across tasks.

[30] Note that, in Frydman and Jin (2022) and Khaw et al. (2021), the probability of the lottery payoff is not only an objective quantity but remains fixed over all trials. Only differences in the payoffs are thus important for choices, which could render their lottery choice task into a number perception task (see (Alós-Ferrer and Garagnani, 2022, 313).

[31] For risk elicitation methods, Holzmeister and Stefan (2021) show that the within-person inconsistency in risk elicitations across different methods is related to the subjectively perceived elicitation complexity, suggesting that the complexity of a setting could potentially guide the impact of overarching concepts.

a directional association with different levels of altruism are less clear (recall that I did not observe any correlation between measures of cognitive noise and altruism per se). Similar to our treatment effect, this implies that e.g., across contexts of varying complexity, differences in cognitive ability could nonetheless lead to systematically different behavior. Recall also that the association between changes in payments in the Treatment group was flatter compared to the Baseline group. This "flatness" (or insensitivity) is at the center of discussions in Enke and Graeber (2023), Enke et al. (2023a) and especially Enke et al. (2024), who establish cognitive uncertainty as the common driver of such inattentive behavior across over 30 experiments in various decision domains. Transporting this argument to the present setting, this speaks in favor of a dampened expression of selfish preferences in the Treatment group. This is a more nuanced angle on the discussion that investigates associations between cognitive ability and economic preferences more generally (Burks et al., 2009; Chapman et al., 2023; Falk et al., 2018; Stango and Zinman, 2023), especially for associations between social preferences and cognitive ability (Hauge et al., 2009; Chen et al., 2013; Ponti and Rodriguez-Lara, 2015), which mostly focus on associations between the level of preferences and cognitive ability thus far. This is related to an ongoing discussion in risk and time preferences and whether cognitive ability is related to choice mistakes "only" or preference per se (Amador-Hidalgo et al., 2021; Olschewski et al., 2023).

(iv) RT and Metacognition The fourth set of results is related to the link between behavior, RT and metacognition. The main result there is the presence of a treatment effect in "metacognitive" measures in the number comparison, yet an absence of such an effect in the altruism domain. Correlational analyses further show that the link between metacognition, RT and choices in *weaker* in the altruism domain compared to the number comparison domain. One possible explanation for this could be that in domains where an objectively correct solution exists, RT and metacognition (which could be formed from a recollection of the latter (Kiani et al., 2014)) are *better calibrated* because a more direct notion of a "correct" solution is available. In turn, the treatment variation, aimed at increasing cognitive difficulty, could have only an effect on metacognitive reasoning with a clear indication of what a "correct" choice is. This does not imply that metacognitive judgments are detached from internal processes (see the discussion in Fleming (2024) for value-based decisions), yet their determinants and consequences are possibly different and generally remain less well understood (Brus et al., 2021). Economic tasks often contain a strong subjective component of what is "correct" and in turn, could imply that the overall link between metacognition and subjective preferences "plays out" differently compared to settings with more clear notions of choice correctness. This points towards a difference between lottery choices (which also remains dependent on subjective preferences) and altruism choices: In the former, a benchmark choice, i.e., the one that maximizes expected value is available. Such a "virtually objective" benchmark is lacking when making altruistic choices. Ultimately, this could imply that the impact of the selection processes described in Enke et al. (2023b) – which depend on the relationship between confidence and performance – may lead to different equilibrium outcomes in settings of pure value-based decisions where no clear notion of choice bias exists.

6 Conclusion

In this paper, I study altruistic choices through the lens of a cognitively noisy decision-maker. I ran an experiment that elicited altruistic choices – choosing between taking an amount $self$ or giving an amount $other$ – featuring a between-subject manipulation of cognitive difficulty of choosing by introducing to-be-calculated sums instead of plain monetary values. I observe both a flatter association between changes in payments and choices in the Treatment group as well as overall more altruistic choices. I repeated the trials of the experiment in a comparable number comparison task, where participants had to judge which of two numbers was larger. In this task, I observe a similar treatment effect, which suggests that the perception of numerical magnitudes – in particular an intuitive "intermediate" perception of numerical values – is responsible for the observed group differences in both tasks. In addition to these treatment differences, I observe (correlational) associations between number comparison behavior, cognitive ability and altruistic choice.

Altruistic preferences – and social preferences more generally – are thus not immune to the cognitive difficulty of their implementation, which further implies that at least part of observed social preferences is due to (individual differences in) cognitive noise, which in turn may be related to cognitive ability. This suggests that the expression of social preferences is likely to be context-dependent if different contexts invoke differences in the "noisiness of perception" or have different complexity. Ultimately, this is an important implication if (revealed) social preferences are used as the basis for welfare calculations.

A caveat of this paper remains in that the treatment effect in the experiment remained relatively small and altruistic behavior did not react that much to increasing noise. However, this could be both related to the fact that altruism choices as operationalized here remained relatively simple and the chosen treatment variation represents a relatively mild increase in cognitive noise. Both, in turn, imply that the observed group differences are likely a *lower bound* on the influence of cognitive noise on social preferences more generally. Other decisions involving social preferences are often much more complex to carry out: Both in more involved laboratory environments, e.g., in choosing between payoff allocations as in the popular binary dictator game and in real-world scenarios featuring social preferences that often involve multiple trade-offs, decisions are likely *more* prone to be affected by cognitive noise. Exploring these effects is therefore a promising avenue for future research.

References

Alós-Ferrer, C. and Garagnani, M. (2022). Strength of preference and decisions under risk. *Journal of Risk and Uncertainty*, 64(3):309–329.

Alós-Ferrer, C., Garagnani, M., and Hügelschäfer, S. (2016). Cognitive Reflection, Decision Biases, and Response Times. *Frontiers in Psychology*, 7.

Amador-Hidalgo, L., Brañas-Garza, P., Espín, A. M., García-Muñoz, T., and Hernández-Román, A. (2021). Cognitive abilities and risk-taking: Errors, not preferences. *European Economic Review*, 134:103694.

Amelio, A. (2022). Cognitive Uncertainty and Overconfidence. *ECONtribute Discussion Paper No. 173*.

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.

Assenza, T., Cardaci, A., and Delli Gatti, D. (2019). Perceived Wealth, Cognitive Sophistication and Behavioral Inattention. *SSRN Electronic Journal*.

Augenblick, N., Lazarus, E., and Thaler, M. (2022). Overinference from Weak Signals and Underinference from Strong Signals. *SSRN Electronic Journal*.

Barretto-García, M., De Hollander, G., Grueschow, M., Polanía, R., Woodford, M., and Ruff, C. C. (2023). Individual risk attitudes arise from noise in neurocognitive magnitude representations. *Nature Human Behaviour*, 7(9):1551–1567.

Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oka, P., Oprescu, M., and Syrgkanis, V. (2019). EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation.

Bavard, S., Stuchlý, E., Konovalov, A., and Gluth, S. (2024). Humans can infer social preferences from decision speed alone. *PLOS Biology*, 22(6):e3002686.

Bech, M., Kjaer, T., and Lauridsen, J. (2011). Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment. *Health Economics*, 20(3):273–286.

Beißert, H., Köhler, M., Rempel, M., and Beierlein, C. (2014). Eine deutschsprachige Kurzskala zur Messung des Konstrukts Need for Cognition. *GESEIS Working Papers*, 32.

Bellemare, C., Kröger, S., and Van Soest, A. (2011). Preferences, intentions, and expectation violations: A large-scale experiment with a representative subject pool. *Journal of Economic Behavior & Organization*, 78(3):349–365.

Bernheim, B. D. and Stark, O. (1988). Altruism within the Family Reconsidered: Do Nice Guys Finish Last? *The American Economic Review*, 78(5):1034–1045.

Betsch, C. (2004). Präferenz für Intuition und Deliberation (PID) - Inventar zur Erfassung von affekt- und kognitionsbasiertem Entscheiden. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 25(4):179–197.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P. A., Horsfall, P., and Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20:28:1–28:6.

Bland, J. R. (2023). Bayesian Model Selection and Prior Calibration for Structural Models in Economic Experiments: Some Guidance for the Practitioner. *SSRN Electronic Journal*.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, 90(1):166–193.

Bruhin, A., Fehr, E., and Schunk, D. (2019). The many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences. *Journal of the European Economic Association*, 17(4):1025–1069.

Brus, J., Aebersold, H., Grueschow, M., and Polania, R. (2021). Sources of confidence in value-based choice. *Nature Communications*, 12(1):7337.

Burks, S. V., Carpenter, J. P., Goette, L., and Rustichini, A. (2009). Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences*, 106(19):7745–7750.

Cacioppo, J. T. and Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1):116–131.

Cantlon, J. F., Libertus, M. E., Pinel, P., Dehaene, S., Brannon, E. M., and Pelphrey, K. A. (2009). The Neural Development of an Abstract Concept of Number. *Journal of Cognitive Neuroscience*, 21(11):2217–2229.

Cappelen, A. W., Nielsen, U. H., Tungodden, B., Tyran, J.-R., and Wengström, E. (2016). Fairness is intuitive. *Experimental Economics*, 19(4):727–740.

Carpenter, J. P. and Robbett, A. (2022). Measuring Socially Appropriate Social Preferences. *SSRN Electronic Journal*.

Chapman, J., Dean, M., Ortoleva, P., Snowberg, E., and Camerer, C. (2023). Econographics. *Journal of Political Economy Microeconomics*, 1(1):115–161.

Charness, G. and Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Chen, C.-C., Chiu, I.-M., Smith, J., and Yamada, T. (2013). Too smart to be selfish? Measures of cognitive ability, social preferences, and consistency. *Journal of Economic Behavior & Organization*, 90:112–122.

Chen, F., Zhu, Z., Shen, Q., Krajbich, I., and Hare, T. A. (2023). Intrachoice Dynamics Shape Social Decisions. *Management Science*.

Chew, S. H., Miao, B., Shen, Q., and Zhong, S. (2022). Multiple-switching behavior in choice-list elicitation of risk preference. *Journal of Economic Theory*, 204:105510.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., and Garcia-Retamero, R. (2012). Measuring Risk Literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, 7(1):25–47.

Cooper, D. J. and Kagel, J. H. (2016). Other-Regarding Preferences A Selective Survey of Experimental Results. In Kagel, J. H. and Roth, A. E., editors, *The Handbook of Experimental Economics, Volume Two*. Princeton University Press, Princeton.

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.

Dehaene, S. (1993). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In Haggard, P., Rossetti, Y., and Kawato, M., editors, *Sensorimotor Foundations of Higher Cognition*, 22, pages 527–574.

Dehaene, S. (2011). *The Number Sense: How the Mind Creates Mathematics, Revised and Updated Edition*. Oxford University Press, USA, New York, updated edition edition.

Dehaene, S. and Marques, J. F. (2002). Cognitive euroscience: Scalar variability in price estimation and the cognitive consequences of switching to the euro. *The Quarterly Journal of Experimental Psychology Section A*, 55(3):705–731.

Diester, I. and Nieder, A. (2007). Semantic Associations between Signs and Numerical Categories in the Prefrontal Cortex. *PLoS Biology*, 5(11):e294.

Doya, K., Ishii, S., Pouget, A., Rao, R. P. N., Sejnowski, T. J., and Poggio, T. A., editors (2006). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. MIT Press.

Echeverry, D., Figueroa, M. C., and Polanıa-Reyes, S. (2023). Structural Identification of Social Preferences: Heterogeneity Matters for Incentives. *University de Nvarra Working Papers*, (2).

Enke, B. (2024). The Cognitive Turn in Behavioral Economics.

Enke, B. and Graeber, T. (2023). Cognitive Uncertainty. *The Quarterly Journal of Economics*, 138(4):2021–2067.

Enke, B., Graeber, T., and Oprea, R. (2023a). Complexity and Hyperbolic Discounting.

Enke, B., Graeber, T., and Oprea, R. (2023b). Confidence, Self-Selection, and Bias in the Aggregate. *American Economic Review*, 113(7):1933–1966.

Enke, B., Graeber, T., Oprea, R., and Yang, J. (2024). Behavioral Attenuation.

Exley, C. and Kessler, J. (2024). Motivated Errors. *American Economic Review*, 114(4):961–987.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global Evidence on Economic Preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.

Falk, A., Becker, A., Dohmen, T., Huffman, D., and Sunde, U. (2023). The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences. *Management Science*, 69(4):1935–1950.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315.

Fehr, E. and Charness, G. (2023). Social Preferences: Fundamental Characteristics and Economic Consequences.

Fehr, E. and Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Fehr, E. and Schmidt, K. M. (2006). Chapter 8 The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories. In *Handbook of the Economics of Giving, Altruism and Reciprocity*, volume 1, pages 615–691. Elsevier.

Feigenson, L., Dehaene, S., and Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, 8(7):307–314.

Fisman, R., Kariv, S., and Markovits, D. (2007). Individual Preferences for Giving. *American Economic Review*, 97(5):1858–1876.

Fleming, S. M. (2024). Metacognition and Confidence: A Review and Synthesis. *Annual Review of Psychology*, (75):241–268.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4):25–42.

Frydman, C. and Jin, L. J. (2022). Efficient Coding and Risky Choice. *The Quarterly Journal of Economics*, 137(1):161–213.

Frydman, C. and Nunnari, S. (2023). Coordination with Cognitive Noise.

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

Gelman, A. (2020). Prior Choice Recommendations. https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2021). Bayesian Data Analysis. (Third Edition).

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543.

Graf, C., Vetschera, R., and Zhang, Y. (2013). Parameters of social preference functions: Measurement and external validity. *Theory and Decision*, 74(3):357–382.

Hauge, K. E., Brekke, K. A., Johansson, O., Johansson, O., and Svedsäter, H. (2009). Are Social Preferences Skin Deep? Dictators under Cognitive Load. *University of Gothenburg Working Papers in Economics*, (No. 371).

Holzmeister, F. and Stefan, M. (2021). The risk elicitation puzzle revisited: Across-methods (in)consistency? *Experimental Economics*, 24(2):593–616.

Hossain, T. and Okui, R. (2013). The Binarized Scoring Rule. *The Review of Economic Studies*, 80(3):984–1001.

Hutcherson, C. A., Bushong, B., and Rangel, A. (2015). A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron*, 87(2):451–462.

Khaw, M. W., Li, Z., and Woodford, M. (2021). Cognitive Imprecision and Small-Stakes Risk Aversion. *The Review of Economic Studies*, 88(4):1979–2013.

Kiani, R., Corthell, L., and Shadlen, M. N. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, 84(6):1329–1342.

Klockmann, V., Von Schenk, A., and Villeval, M. C. (2022). Artificial intelligence, ethics, and intergenerational responsibility. *Journal of Economic Behavior & Organization*, 203:284–317.

König-Kersting, C. (2021). On the Robustness of Social Norm Elicitation. *Working Papers in Economics and Statistics - University of Innsbruck*.

Krajbich, I., Bartling, B., Hare, T., and Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, 6(1):7455.

Krupka, E. L. and Weber, R. A. (2013). Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? *Journal of the European Economic Association*, 11(3):495–524.

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, Boston, 2 edition.

Kumar, R., Carroll, C., Hartikainen, A., and Martin, O. (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, 4(33):1143.

Laurent, G. and Vanhuele, M. (2023). How Do Consumers Read and Encode a Price? *Journal of Consumer Research*, 50(3):510–532.

Lepper, M. (2024). Excuse-Based Procrastination. *mimeo*.

Levine, D. K. (1998). Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics*, 1(3):593–622.

Longo, M. R. and Lourenco, S. F. (2007). Spatial attention and the mental number line: Evidence for characteristic biases and compression. *Neuropsychologia*, 45(7):1400–1407.

Luce, R. D. (1991). *Response Times: Their Role in Inferring Elementary Mental Organization*. Number 8 in Oxford Psychology Series. Univ. Press, Oxford, 1. issued as paperback edition.

McFadden, D. (1981). Econometric Models for Probabilistic Choice. In Manski, C. and McFadden, D., editors, *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press, Cambridge, MA.

Merkel, A. L. and Lohse, J. (2019). Is fairness intuitive? An experiment accounting for subjective utility differences under time pressure. *Experimental Economics*, 22(1):24–50.

Moyer, R. S. and Landauer, T. K. (1967). Time required for Judgements of Numerical Inequality. *Nature*, 215(5109):1519–1520.

Nieder, A. and Dehaene, S. (2009). Representation of Number in the Brain. *Annual Review of Neuroscience*, 32(1):185–208.

Nunnari, S. and Pozzi, M. (2022). Meta-Analysis of Distributional Preferences Estimates.

Olschewski, S., Rieskamp, J., and Hertwig, R. (2023). The link between cognitive abilities and risk preference depends on measurement. *Scientific Reports*, 13(1):21151.

Olschewski, S. and Scheibehenne, B. (2024). What's in a sample? Epistemic uncertainty and metacognitive awareness in risk taking. *Cognitive Psychology*, 149:101642.

Oprea, R. (2024). Decisions Under Risk are Decisions Under Complexity. *American Economic Review*.

Oprea, R. and Vieider, F. (2024). Minding the Gap: On the Origins of Probability Weighting and the Description-Experience Gap.

Özdemir, S., Mohamed, A. F., Johnson, F. R., and Hauber, A. B. (2010). Who pays attention in stated-choice surveys? *Health Economics*, 19(1):111–118.

Phan, D., Pradhan, N., and Jankowiak, M. (2019). Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv preprint arXiv:1912.11554*.

Pins, D. and Bonnet, C. (1996). On the relation between stimulus intensity and processing time: Piéron's law and choice reaction time. *Perception & Psychophysics*, 58(3):390–400.

Poggi, L. (2021). Learning dynamics in optimal decision making. *mimeo*.

Polanía, R., Woodford, M., and Ruff, C. C. (2019). Efficient coding of subjective value. *Nature Neuroscience*, 22(1):134–142.

Ponti, G. and Rodriguez-Lara, I. (2015). Social preferences and cognitive reflection: Evidence from a dictator game experiment. *Frontiers in Behavioral Neuroscience*, 9.

Prat-Carrabin, A. and Woodford, M. (2022). Efficient coding of numbers explains decision bias and noise. *Nature Human Behaviour*, 6(8):1142–1152.

Pustejovsky, J. E. and Tipton, E. (2018). Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models. *Journal of Business & Economic Statistics*, 36(4):672–683.

Radbruch, L., Sabatowski, R., Elsner, F., Everts, J., Mendoza, T., and Cleeland, C. (2003). Validation of the German Version of the Brief Fatigue Inventory. *Journal of Pain and Symptom Management*, 25(5):449–458.

Rand, D. G., Greene, J. D., and Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416):427–430.

Rand, D. G. and Kraft-Todd, G. T. (2014). Reflection does not undermine self-interested prosociality. *Frontiers in Behavioral Neuroscience*, 8.

Schlag, K. and Van Der Weele, J. (2013). Eliciting Probabilities, Means, Medians, Variances and Covariances without Assuming Risk Neutrality. *Theoretical Economics Letters*, 3(1):38–42.

Schunk, D. and Betsch, C. (2006). Explaining heterogeneity in utility functions by individual differences in decision modes. *Journal of Economic Psychology*, 27(3):386–401.

Schwappach, D. L. and Strasmann, T. J. (2006). "Quick and dirty numbers"?: The reliability of a stated-preference technique for the measurement of preferences for resource allocation. *Journal of Health Economics*, 25(3):432–448.

Stango, V. and Zinman, J. (2023). We Are All Behavioural, More, or Less: A Taxonomy of Consumer Decision-Making. *The Review of Economic Studies*, 90(3):1470–1498.

Stocker, A. A. and Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–585.

Toplak, M. E., West, R. F., and Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, 20(2):147–168.

van Leeuwen, B. and Alger, I. (2023). Estimating Social Preferences and Kantian Morality in Strategic Interactions.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. *Bayesian Analysis*, 16(2).

Vieider, F. (2022). Noisy coding of time and reward discounting.

Vieider, F. (2024a). Bayesian Estimation of Decision Models. https://fvieider.quarto.pub/bstats/.

Vieider, F. (2024b). Decisions Under Uncertainty as Bayesian Inference on Choice Options. *Management Science*, page mnsc.2023.00265.

Von Schenk, A., Klockmann, V., and Köbis, N. (2023). Social Preferences Toward Humans and Machines: A Systematic Experiment on the Role of Machine Payoffs. *Perspectives on Psychological Science*.

Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Watanabe, S. (2013). A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research*, 14:867–897.

Wei, X.-X. and Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, 114(38):10244–10249.

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604.

Woodford, M. (2012). Prospect Theory as Efficient Perceptual Distortion. *American Economic Review*, 102(3):41–46.

Woodford, M. (2020). Modeling Imprecision in Perception, Valuation, and Choice. *Annual Review of Economics*, 12(1):579–601.

Xie, Y. and Carlin, B. P. (2006). Measures of Bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, 136(10):3458–3477.

A Appendix

A.1 Theory

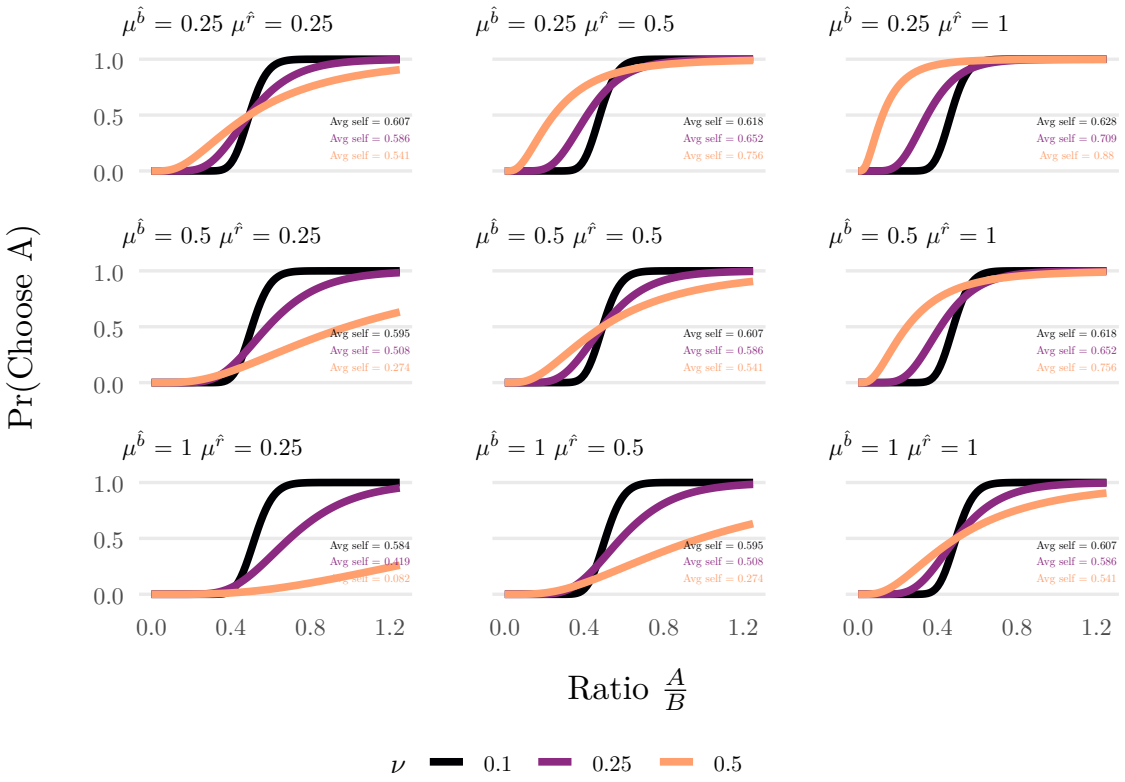


Figure A.1: Impact of Noise ν on Altruistic Choices (Same Prior Standard Deviation). This shows the impact of changes in cognitive noise ν on the choice function 7 depending on different values of $\mu^{\hat{b}}$ and $\mu^{\hat{r}}$. Note that $\sigma^{\hat{b}} = \sigma^{\hat{r}} = 0.5$ and $\beta = 0.33$ in all panels.

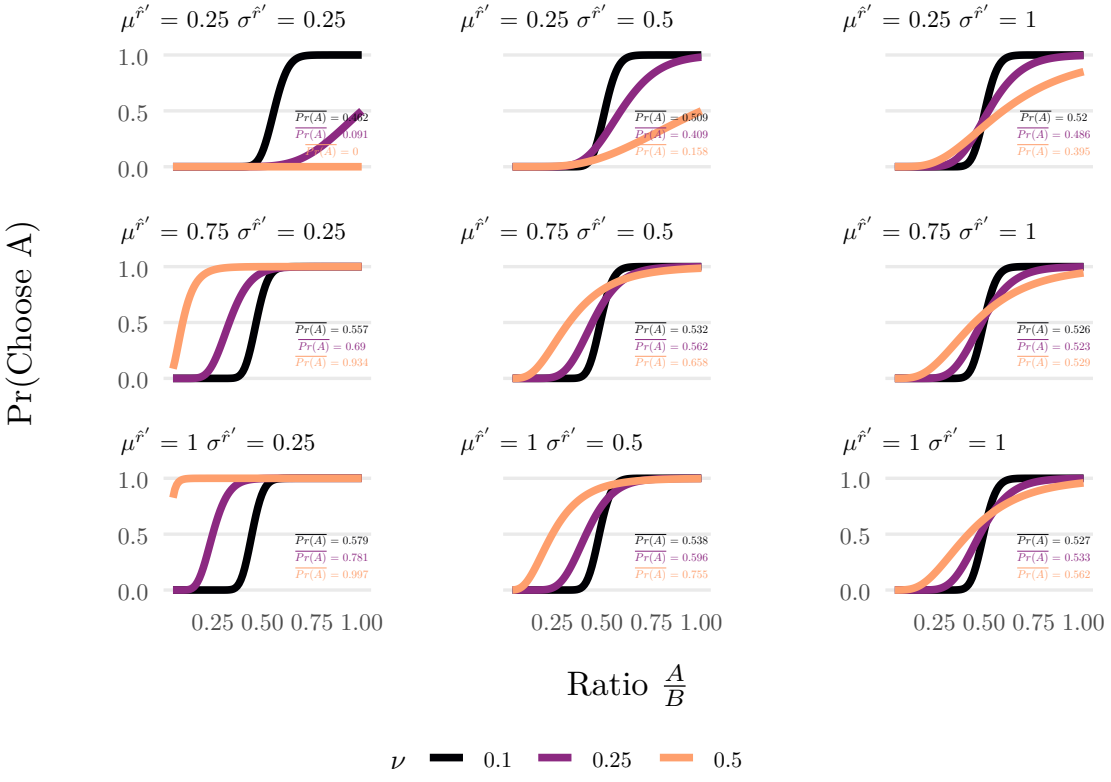


Figure A.2: Impact of Noise ν on Number Comparison. This shows the impact of changes in cognitive noise ν on the choice function 9 depending on different values of $r^{'}$ and $\sigma^{\hat{r'}}$.

A.2 Graphs and Figures Experiment

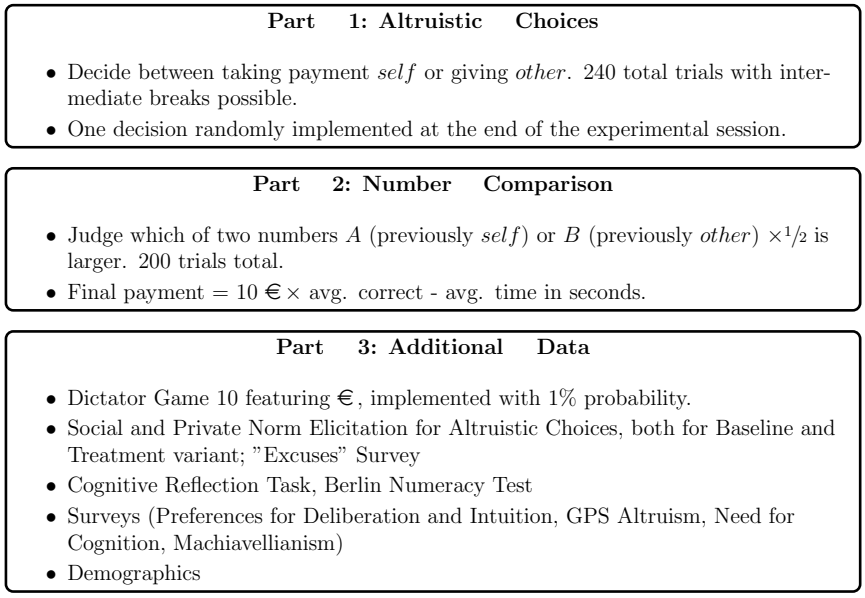


Figure A.3: Graphical Outline of an Experimental Session.

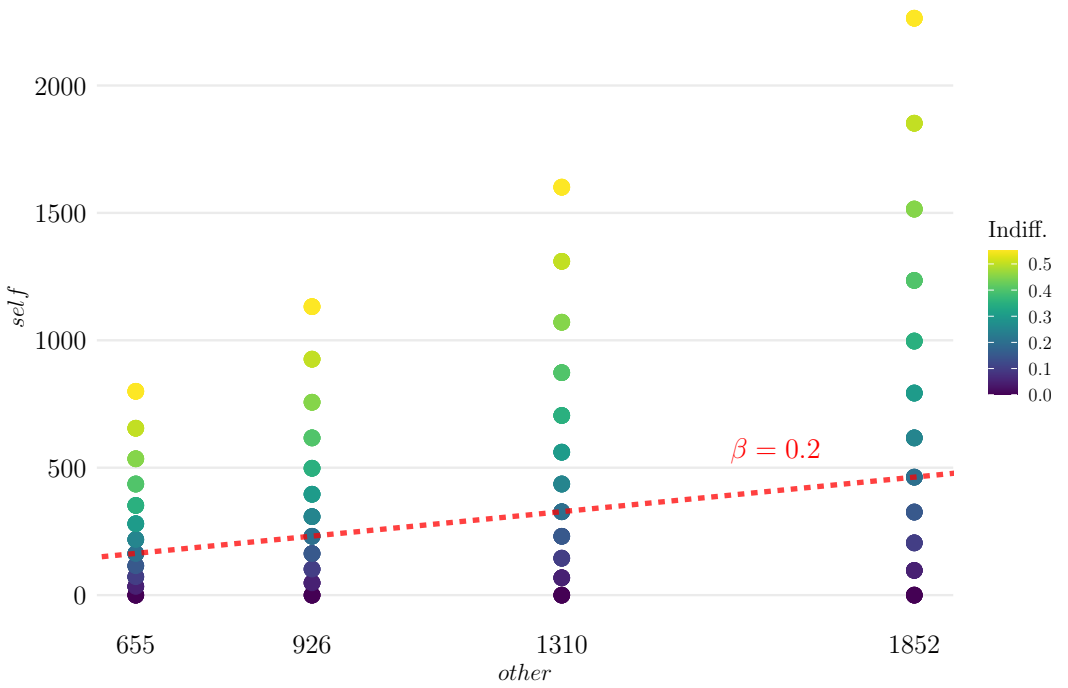


Figure A.4: Payment Combinations in Altruistic Choice Task (in Eurocents). This graph shows the 48 unique combinations of *self* and *other* in the experiment trials. Each combination is repeated five times, totaling 240 decisions, with one decisions randomly implemented. Note that I instructed participants precisely like this, but not each trial had the same chance of being drawn: Instead of drawing from a uniform distribution across trials, I overweighted trials of smaller stakes (i.e., where the sum of *self* and *other* is small) to be more likely to be drawn. Details and implementation are available upon request. The indifference threshold for a noiseless decision maker with a $\beta = 0.2$ is drawn for illustration purposes. This DM will always decide for *other* in the trials below and for *self* above this line

Table A.1: Overview of 240 trials of Altruism Task

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
0	1	655	0	0	0	543	112
0	1	655	0	0	0	629	26
0	1	655	0	0	0	490	165
0	1	655	0	0	0	32	623
0	1	655	0	0	0	540	115
1	1	655	34	14	20	638	17
1	1	655	34	23	11	20	635
1	1	655	34	22	12	643	12
1	1	655	34	15	19	14	641
1	1	655	34	22	12	10	645
2	1	655	72	24	48	627	28
2	1	655	72	35	37	34	621
2	1	655	72	58	14	26	629
2	1	655	72	61	11	40	615

Table A.1: Overview of 240 trials of Altruism Task *(continued)*

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
2	1	655	72	37	35	35	620
3	1	655	115	31	84	33	622
3	1	655	115	36	79	621	34
3	1	655	115	39	76	617	38
3	1	655	115	28	87	645	10
3	1	655	115	34	81	585	70
4	1	655	163	62	101	11	644
4	1	655	163	12	151	10	645
4	1	655	163	15	148	643	12
4	1	655	163	13	150	643	12
4	1	655	163	140	23	554	101
5	1	655	218	167	51	47	608
5	1	655	218	164	54	509	146
5	1	655	218	172	46	622	33
5	1	655	218	81	137	41	614
5	1	655	218	18	200	122	533
6	1	655	280	170	110	588	67
6	1	655	280	234	46	60	595
6	1	655	280	90	190	104	551
6	1	655	280	161	119	506	149
6	1	655	280	126	154	15	640
7	1	655	352	107	245	557	98
7	1	655	352	161	191	68	587
7	1	655	352	227	125	583	72
7	1	655	352	310	42	645	10
7	1	655	352	170	182	486	169
8	1	655	436	68	368	51	604
8	1	655	436	331	105	97	558
8	1	655	436	425	11	485	170
8	1	655	436	326	110	634	21
8	1	655	436	312	124	158	497
9	1	655	535	199	336	471	184
9	1	655	535	413	122	71	584
9	1	655	535	398	137	27	628
9	1	655	535	426	109	478	177
9	1	655	535	222	313	443	212
10	1	655	655	253	402	79	576
10	1	655	655	277	378	82	573
10	1	655	655	332	323	575	80
10	1	655	655	361	294	565	90

Table A.1: Overview of 240 trials of Altruism Task *(continued)*

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
10	1	655	655	156	499	419	236
11	1	655	800	740	60	515	140
11	1	655	800	678	122	260	395
11	1	655	800	503	297	635	20
11	1	655	800	311	489	39	616
11	1	655	800	744	56	244	411
12	2	926	0	0	0	850	76
12	2	926	0	0	0	836	90
12	2	926	0	0	0	254	672
12	2	926	0	0	0	418	508
12	2	926	0	0	0	391	535
13	2	926	48	13	35	10	916
13	2	926	48	11	37	10	916
13	2	926	48	36	12	19	907
13	2	926	48	32	16	26	900
13	2	926	48	30	18	914	12
14	2	926	102	84	18	11	915
14	2	926	102	39	63	32	894
14	2	926	102	76	26	909	17
14	2	926	102	35	67	19	907
14	2	926	102	66	36	20	906
15	2	926	163	70	93	69	857
15	2	926	163	138	25	28	898
15	2	926	163	111	52	48	878
15	2	926	163	39	124	114	812
15	2	926	163	25	138	17	909
16	2	926	231	214	17	16	910
16	2	926	231	62	169	53	873
16	2	926	231	35	196	914	12
16	2	926	231	67	164	864	62
16	2	926	231	161	70	126	800
17	2	926	308	90	218	45	881
17	2	926	308	101	207	169	757
17	2	926	308	22	286	905	21
17	2	926	308	84	224	38	888
17	2	926	308	173	135	99	827
18	2	926	396	11	385	10	916
18	2	926	396	184	212	733	193
18	2	926	396	74	322	45	881
18	2	926	396	170	226	896	30

Table A.1: Overview of 240 trials of Altruism Task *(continued)*

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
18	2	926	396	325	71	187	739
19	2	926	498	63	435	51	875
19	2	926	498	264	234	879	47
19	2	926	498	330	168	273	653
19	2	926	498	325	173	779	147
19	2	926	498	366	132	50	876
20	2	926	617	486	131	200	726
20	2	926	617	410	207	148	778
20	2	926	617	409	208	775	151
20	2	926	617	416	201	186	740
20	2	926	617	106	511	470	456
21	2	926	757	565	192	171	755
21	2	926	757	604	153	106	820
21	2	926	757	480	277	152	774
21	2	926	757	557	200	821	105
21	2	926	757	733	24	470	456
22	2	926	926	224	702	647	279
22	2	926	926	79	847	730	196
22	2	926	926	117	809	83	843
22	2	926	926	46	880	727	199
22	2	926	926	370	556	567	359
23	2	926	1132	827	305	711	215
23	2	926	1132	669	463	146	780
23	2	926	1132	1072	60	863	63
23	2	926	1132	598	534	222	704
23	2	926	1132	867	265	323	603
24	3	1310	0	0	0	963	347
24	3	1310	0	0	0	898	412
24	3	1310	0	0	0	726	584
24	3	1310	0	0	0	876	434
24	3	1310	0	0	0	459	851
25	3	1310	68	35	33	1288	22
25	3	1310	68	28	40	1294	16
25	3	1310	68	31	37	26	1284
25	3	1310	68	26	42	10	1300
25	3	1310	68	29	39	1283	27
26	3	1310	145	121	24	1288	22
26	3	1310	145	42	103	1293	17
26	3	1310	145	79	66	1286	24
26	3	1310	145	121	24	1212	98

Table A.1: Overview of 240 trials of Altruism Task *(continued)*

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
26	3	1310	145	14	131	1298	12
27	3	1310	231	80	151	1292	18
27	3	1310	231	140	91	10	1300
27	3	1310	231	34	197	110	1200
27	3	1310	231	35	196	28	1282
27	3	1310	231	127	104	74	1236
28	3	1310	327	182	145	1238	72
28	3	1310	327	167	160	1259	51
28	3	1310	327	62	265	1257	53
28	3	1310	327	242	85	1154	156
28	3	1310	327	79	248	1124	186
29	3	1310	436	346	90	1293	17
29	3	1310	436	261	175	153	1157
29	3	1310	436	112	324	1233	77
29	3	1310	436	143	293	215	1095
29	3	1310	436	145	291	72	1238
30	3	1310	561	298	263	1040	270
30	3	1310	561	550	11	10	1300
30	3	1310	561	202	359	1254	56
30	3	1310	561	515	46	934	376
30	3	1310	561	470	91	884	426
31	3	1310	705	655	50	1273	37
31	3	1310	705	442	263	1161	149
31	3	1310	705	665	40	278	1032
31	3	1310	705	627	78	21	1289
31	3	1310	705	597	108	10	1300
32	3	1310	873	849	24	521	789
32	3	1310	873	704	169	1292	18
32	3	1310	873	758	115	1271	39
32	3	1310	873	395	478	62	1248
32	3	1310	873	832	41	783	527
33	3	1310	1071	512	559	416	894
33	3	1310	1071	246	825	217	1093
33	3	1310	1071	942	129	72	1238
33	3	1310	1071	493	578	1262	48
33	3	1310	1071	718	353	656	654
34	3	1310	1310	108	1202	814	496
34	3	1310	1310	63	1247	1149	161
34	3	1310	1310	750	560	219	1091
34	3	1310	1310	944	366	235	1075

Table A.1: Overview of 240 trials of Altruism Task *(continued)*

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
34	3	1310	1310	304	1006	254	1056
35	3	1310	1601	1045	556	575	735
35	3	1310	1601	1005	596	637	673
35	3	1310	1601	1465	136	800	510
35	3	1310	1601	1459	142	667	643
35	3	1310	1601	1134	467	587	723
36	4	1852	0	0	0	622	1230
36	4	1852	0	0	0	1453	399
36	4	1852	0	0	0	1022	830
36	4	1852	0	0	0	1110	742
36	4	1852	0	0	0	734	1118
37	4	1852	97	85	12	10	1842
37	4	1852	97	34	63	27	1825
37	4	1852	97	43	54	41	1811
37	4	1852	97	27	70	1828	24
37	4	1852	97	37	60	54	1798
38	4	1852	205	54	151	1740	112
38	4	1852	205	91	114	1777	75
38	4	1852	205	57	148	1818	34
38	4	1852	205	83	122	1785	67
38	4	1852	205	122	83	113	1739
39	4	1852	326	216	110	24	1828
39	4	1852	326	109	217	198	1654
39	4	1852	326	45	281	198	1654
39	4	1852	326	64	262	239	1613
39	4	1852	326	222	104	81	1771
40	4	1852	463	74	389	188	1664
40	4	1852	463	293	170	77	1775
40	4	1852	463	385	78	1821	31
40	4	1852	463	150	313	1706	146
40	4	1852	463	259	204	171	1681
41	4	1852	617	367	250	61	1791
41	4	1852	617	331	286	1708	144
41	4	1852	617	85	532	1817	35
41	4	1852	617	414	203	1838	14
41	4	1852	617	302	315	277	1575
42	4	1852	793	50	743	1358	494
42	4	1852	793	649	144	42	1810
42	4	1852	793	698	95	66	1786
42	4	1852	793	431	362	348	1504

Table A.1: Overview of 240 trials of Altruism Task *(continued)*

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
42	4	1852	793	245	548	142	1710
43	4	1852	997	239	758	1188	664
43	4	1852	997	964	33	960	892
43	4	1852	997	768	229	1774	78
43	4	1852	997	374	623	330	1522
43	4	1852	997	772	225	1442	410
44	4	1852	1235	1069	166	1318	534
44	4	1852	1235	1055	180	759	1093
44	4	1852	1235	103	1132	1781	71
44	4	1852	1235	715	520	1516	336
44	4	1852	1235	307	928	1636	216
45	4	1852	1515	1276	239	1119	733
45	4	1852	1515	1276	239	118	1734
45	4	1852	1515	1237	278	921	931
45	4	1852	1515	1165	350	749	1103
45	4	1852	1515	928	587	1768	84
46	4	1852	1852	566	1286	1161	691
46	4	1852	1852	536	1316	1809	43
46	4	1852	1852	1502	350	441	1411
46	4	1852	1852	454	1398	1820	32
46	4	1852	1852	773	1079	1455	397
47	4	1852	2264	655	1609	1022	830
47	4	1852	2264	1457	807	1310	542
47	4	1852	2264	1287	977	530	1322
47	4	1852	2264	100	2164	176	1676
47	4	1852	2264	107	2157	1339	513

A.3 Additional Results

Regressions

	(1)	(2)	(3)	(4)
Treatment Group	-0.022***	-0.022	0.009	-0.014
	(0.004)	(0.024)	(0.030)	(0.023)
Ratio self/other			0.900***	0.867***
			(0.018)	(0.013)
Treatment Group * Ratio			-0.067**	
			(0.026)	
Intercept	0.452***	0.452***	0.032	0.052***
	(0.003)	(0.016)	(0.021)	(0.019)
Random Effects	No	No	No	Yes
Clustered Standard Errors	No	Yes	Yes	Yes
Num.Obs.	72000	72000	72000	72000
Unique Obs	300	300	300	300
R2	0.001	0.001	0.431	0.517

Table A.2: Altruism Task Treatment Effect Regression. Linear probability model with choice for *self* as dependent variable. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky and Tipton, 2018)) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

	(1)	(2)	(3)	(4)
Treatment Group	-0.057***	-0.057	0.154	0.663***
	(0.009)	(0.060)	(0.137)	(0.171)
Ratio self/other			3.380***	6.225***
			(0.153)	(0.065)
Treatment Group * Ratio			-0.566***	-1.757***
			(0.197)	(0.078)
Intercept	-0.121***	-0.121***	-1.654***	-3.056***
	(0.007)	(0.041)	(0.100)	(0.122)
Random Effects	No	No	No	Yes
Clustered Standard Errors	No	Yes	Yes	No
Unique Obs	300	300	300	300
pseudo R2	0	0	0.375	-
Num.Obs.	72000	72000	72000	72000

Table A.3: Altruism Task Treatment Effect Probit Model. Probit Model with choice for *self* as dependent variable. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky and Tipton, 2018)) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

	(1)	(2)	(3)
Treatment Group	-0.037***	-0.220***	-0.220***
	(0.005)	(0.022)	(0.022)
Ratio A/B		-0.932***	-0.932***
		(0.005)	(0.005)
Treatment Group * Ratio		0.114***	0.114***
		(0.011)	(0.011)
Intercept	0.388***	1.880***	1.880***
	(0.002)	(0.009)	(0.009)
Random Effects	No	No	Yes
Clustered Standard Erorrs	Yes	Yes	Yes
Num.Obs.	60000	60000	60000
Unique Obs	300	300	300
R2	0.001	0.794	0.798

Table A.4: Number Comparison Treatment Effect Regression. Linear probability model with choice for A as dependent variable. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky and Tipton, 2018)) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

	(1)	(2)	(3)	(4)
Treatment Group	-0.099***	-0.099***	0.839***	0.800***
	(0.010)	(0.012)	(0.286)	(0.077)
Ratio A/B			8.693***	9.131***
			(0.576)	(0.104)
Treatment Group * Ratio			-2.298***	-2.259***
			(0.639)	(0.126)
Intercept	-0.284***	-0.284***	-4.431***	-4.643***
	(0.007)	(0.005)	(0.261)	(0.060)
Random Effects	No	No	No	Yes
Clustered Standard Errors	No	Yes	Yes	No
Unique Obs	300	300	300	300
pseudo R2	0.001	0.001	0.696	-
Num.Obs.	60000	60000	60000	60000

Table A.5: Number Comparison Treatment Effect Probit Model. Probit model with choice for A as dependent variable. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky and Tipton, 2018)) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

	(1)	(2)	(3)
Altruism Task	0.023***	0.023***	0.037***
	(0.002)	(0.002)	(0.002)
Intercept	0.079***	0.073***	-0.015
	(0.002)	(0.020)	(0.021)
Num.Obs.	26400	26400	26400
R2	0.004	0.070	0.191
Participant FE	No	Yes	Yes
Game FE	No	No	Yes
Clustered Standard Errors	No	Yes	Yes

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

Table A.6: Inconsistencies across Tasks Regression. Linear Probability Model. Dependent variable is the standard deviation in a particular game. Clustered standard errors (participant-level) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

A.3.1 Probabilistic Model

Priors

$$\begin{aligned}
\mu^{\nu} &\sim \mathcal{N}(0, 0.25) & \sigma^{\nu} &\sim \text{Exp}(5) \\
\mu^{\frac{\beta}{1-\beta}} &\sim \mathcal{N}(-1, 0.25) & \sigma^{\frac{\beta}{1-\beta}} &\sim \text{Exp}(2) \\
\mu^{\mu^{\hat{r}}} &\sim \mathcal{N}(-0.25, 0.25) & \sigma^{\mu^{\hat{r}}} &\sim \text{Exp}(5) \\
\mu^{\sigma^{\hat{r}}} &\sim \mathcal{N}(0, 0.25) & \sigma^{\sigma^{\hat{r}}} &\sim \text{Exp}(5) \\
\mu^{\mu^{\hat{b}}} &\sim \mathcal{N}(-0.25, 0.25) & \sigma^{\mu^{\hat{b}}} &\sim \text{Exp}(5) \\
\mu^{\sigma^{\hat{b}}} &\sim \mathcal{N}(0, 0.25) & \sigma^{\sigma^{\hat{b}}} &\sim \text{Exp}(5)
\end{aligned}$$

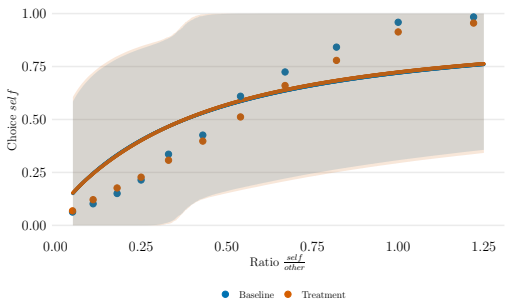
$$\mu^{\nu}_{B} - \mu^{\nu}_{T} \sim \mathcal{N}(0, 0.25)$$

$$\Omega \sim \text{LKJ}(2)$$

Prior Summary and Predictive Checks

	mean	median	hdi 2.5%	hdi 97.5%
Individual Parameters (avg.):				
Altr. Preference β_i	0.287	0.271	0.036	0.543
Noise ν_i	1.074	0.985	0.348	1.916
Prior Std. Outcomes $\sigma_i^{\hat{r}}$	1.06	0.991	0.418	1.776
Prior Std. Preference $\sigma_i^{\hat{b}}$	1.083	0.995	0.381	1.912
Prior Mean Outcomes $\mu_i^{\hat{r}}$	0.822	0.765	0.31	1.408
Prior Mean Altr. Preference $\mu_i^{\hat{b}}$	0.838	0.781	0.32	1.437
Additional Parameter Information (avg.):				
$\sigma_i^{\hat{r}} - \sigma_i^{\hat{b}}$	-0.022	-0.016	-1.184	1.071
$\mu_i^{\hat{r}} - \mu_i^{\hat{b}}$	-0.016	-0.006	-0.86	0.858
$\mu_T^{\nu'} - \mu_B^{\nu'}$	-0.012	-0.006	-0.493	0.457

(a) Prior parameter summary



(b) Average and prior-predicted choices

Figure A.5: Prior summary Probabilistic Model Altruistic Choices. **(a)** Prior parameter values of equation 7 based on 10000 prior samples (i.e., before providing experimental data). Individual parameters are averages (over subjects). Mean and median refer to the mean and median of the prior distribution samples. HDI 2.5% and HDI 97.5% indicate the borders of the 95% highest-density interval (HDI). **(b)** Average and prior-predicted choices, including 95% HDI intervals.

Posterior Summaries

	mean	median	hdi 2.5%	hdi 97.5%	$\hat{R}$
$\mu_T^{\nu} - \mu_B^{\nu}$	0.198	0.198	0.074	0.332	1.00
μ^{ν}	-1.402	-1.403	-1.537	-1.268	1.00
$\mu^{\mu^{\hat{r}}}$	-0.152	-0.156	-0.578	0.306	1.01
$\mu^{\sigma^{\hat{b}}}$	-0.525	-0.536	-0.937	-0.074	1.01
$\mu^{\sigma^{\hat{r}}}$	-0.193	-0.201	-0.617	0.263	1.01
$\mu^{\mu^{\hat{b}}}$	-0.634	-0.644	-1.092	-0.158	1.01
$\mu^{\frac{\beta}{1-\beta}}$	-0.682	-0.681	-0.840	-0.520	1.01
σ^{ν}	0.477	0.474	0.352	0.600	1.00
$\sigma^{\frac{\beta}{1-\beta}}$	0.115	0.113	0.058	0.177	1.01
$\sigma^{\sigma^{\hat{b}}}$	0.443	0.392	0.042	0.966	1.01
$\sigma^{\sigma^{\hat{r}}}$	0.313	0.264	0.002	0.754	1.01
$\sigma^{\mu^{\hat{b}}}$	0.709	0.668	0.068	1.387	1.01
$\sigma^{\mu^{\hat{r}}}$	0.261	0.186	0.001	0.765	1.03

Table A.7: Posterior Parameter Summary Average Individual Parameters Altruistic Choice

	mean	median	hdi 2.5%	hdi 97.5%	$\hat{R}$
Individual Parameters (avg.):					
Altr. Preference β_i	0.342	0.341	0.272	0.415	1
Noise ν_i	0.336	0.323	0.186	0.515	1
Prior Std. Outcomes $\sigma_i^{\hat{r}}$	1.019	0.809	0.173	2.34	1.001
Prior Std. Preference $\sigma_i^{\hat{b}}$	0.772	0.591	0.084	1.885	1.001
Prior Mean Outcomes $\mu_i^{\hat{r}}$	1.023	0.862	0.186	2.19	1
Prior Mean Altr. Preference $\mu_i^{\hat{b}}$	0.797	0.585	0.071	2.017	1.002
Weight on Payments α_i	0.812	0.822	0.648	0.955	1
Prior Threshold δ_i	0.852	0.846	0.559	1.139	1.001
Weight on Preferences γ_i	0.717	0.745	0.372	0.981	1.001
Additional Parameter Information (avg.):					
Noise Baseline ν_{iB}	0.3	0.289	0.171	0.453	1
Noise Treatment ν_{iT}	0.372	0.357	0.2	0.577	1
$\sigma_i^{\hat{r}} - \sigma_i^{\hat{b}}$	0.246	0.211	-1.87	2.452	1.003
$\mu_i^{\hat{r}} - \mu_i^{\hat{b}}$	0.226	0.258	-1.577	2.033	1

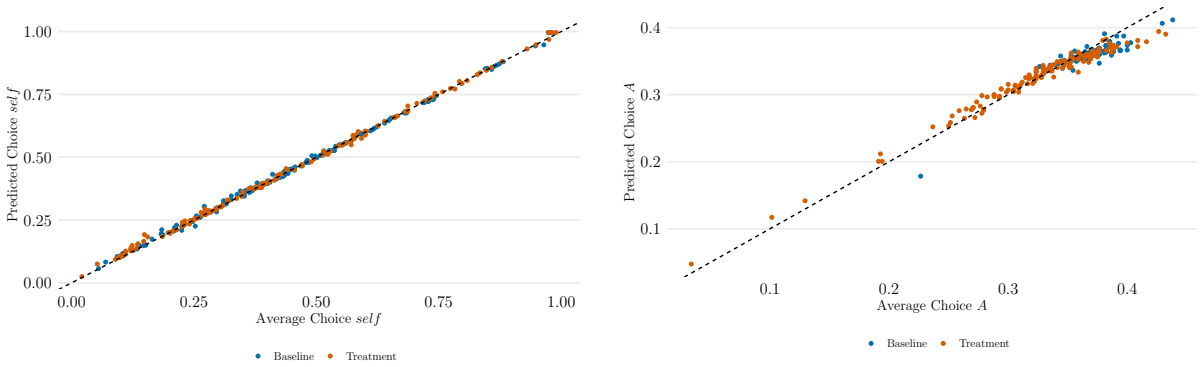
Table A.8: Posterior Parameter Summary Hyper-parameters Altruistic Choice

	mean	median	hdi 2.5%	hdi 97.5%	$\hat{R}$
$\mu^{\nu'}$	-1.739	-1.739	-1.832	-1.645	1.00
$\mu^{\mu^{\hat{r}'}}$	-0.822	-0.817	-0.962	-0.695	1.00
$\mu^{\sigma^{\hat{r}'}}$	-0.269	-0.260	-0.583	0.021	1.03
$\mu_T^{\nu'} - \mu_B^{\nu'}$	0.348	0.349	0.200	0.484	1.00
$\sigma^{\nu'}$	0.269	0.268	0.218	0.321	1.00
$\sigma^{\sigma^{\hat{r}'}}$	0.045	0.040	0.005	0.094	1.03
$\sigma^{\mu^{\hat{r}'}}$	0.350	0.304	0.044	0.776	1.02

Table A.9: Posterior Parameter Summary Hyper-parameters Number Comparison

	mean	median	hdi 2.5%	hdi 97.5%	$\hat{R}$
Individual Parameters (avg.):					
Noise ν_i'	0.233	0.23	0.172	0.301	1
Prior Std. Outcomes $\sigma_i^{\hat{r}'}$	0.785	0.762	0.455	1.153	1.014
Prior Mean Outcomes $\mu_i^{\hat{r}'}$	0.518	0.46	0.157	0.994	1.002
Weight on Payments α_i'	0.905	0.911	0.827	0.969	1.011
Prior Threshold δ_i'	1.103	1.098	1.021	1.193	1.003
Additional Parameter Information (avg.):					
Noise Baseline $\nu_{iB'}$	0.197	0.194	0.144	0.255	1
Noise Treatment $\nu_{iT'}$	0.27	0.266	0.199	0.347	1

Table A.10: Posterior Parameter Summary Hyper-parameters Number Comparison



(a) Altruism Average and Individually Predicted

(b) Number Comparison Average and Individually Predicted

Figure A.6: Individual Average and Predicted Behavior

Model Comparisons

Model	Choice Function	$ELPD_{WAIC}$
Full Model	$Pr(self) = \Phi\frac{\alpha \log \frac{self}{other} - \gamma \log \frac{\beta}{1-\beta} - \log \delta}{\nu \times \sqrt{\alpha^2+\gamma^2}}$	-14,919.78
Same Prior Variance η^2	$Pr(self) = \Phi\frac{\frac{\eta^2}{\eta^2+\nu^2}(\log \frac{self}{other} - \log \frac{\beta}{1-\beta}) - \frac{\nu^2}{\eta^2+\nu^2} \log \frac{\mu^b}{\mu^a}}{\nu \times \sqrt{2\eta}}$	-14,922.64
No Noise in $\frac{\beta}{1-\beta}$	$Pr(self) = \Phi\frac{\alpha \log \frac{self}{other} - \log \frac{\beta}{1-\beta} - \log \frac{1}{\mu^a 1-\alpha}}{\nu \times \alpha}$	-14,923.03
Same Prior Mean κ	$Pr(self) = \Phi\frac{\alpha \log \frac{self}{other} - \gamma \log \frac{\beta}{1-\beta} - \log \kappa^{\alpha-\gamma}}{\nu \times \sqrt{\alpha^2+\gamma^2}}$	-14,923.67
Same Mean and Variance (Pure Noise)	$Pr(self) = \Phi\frac{\log \frac{self}{other} - \log \frac{\beta}{1-\beta}}{\frac{\nu \times \sqrt{2}}{e^\sigma (1-\beta) self + e^\sigma \beta other}}$	-14,935.54
Random Utility	$Pr(self) = \frac{e^\sigma (1-\beta) self}{e^\sigma (1-\beta) self + e^\sigma \beta other}$	-15,656.30

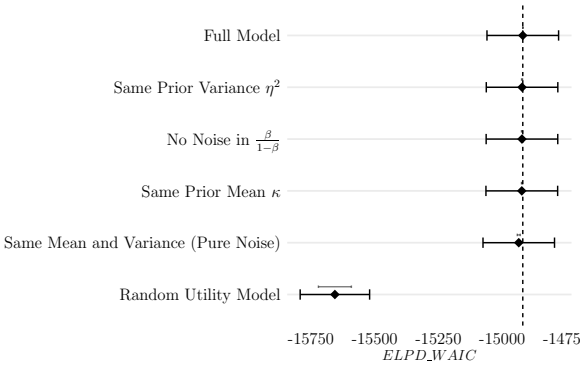


Table A.11 & Figure A.7: Model Comparison Altruistic Choices. $ELPD_{WAIC}$ refers to the expected log predictive density as based on the widely-applicable information criterion (WAIC); A larger $ELPD_{WAIC}$ indicates better model fit. Error bars show the standard error of the respective $ELPD_{WAIC}$ value and the standard error of the $\Delta ELPD_{WAIC}$ value, the $ELPD_{WAIC}$ difference to the best model. Model comparison done via the `arviz`-package (Kumar et al., 2019).

Model	Choice Function	$ELPD_{WAIC}$
Noisy Numbers	$Pr(A) = \Phi\frac{\alpha' \log \frac{A}{B} - \log \frac{1}{2} - \log \delta'}{\nu' \times \alpha'}$	-10463.18
Pure Noise	$Pr(A) = \Phi\frac{\log \frac{A}{B} - \log \frac{1}{2}}{\nu' \times \sqrt{2}}$	-10920.65

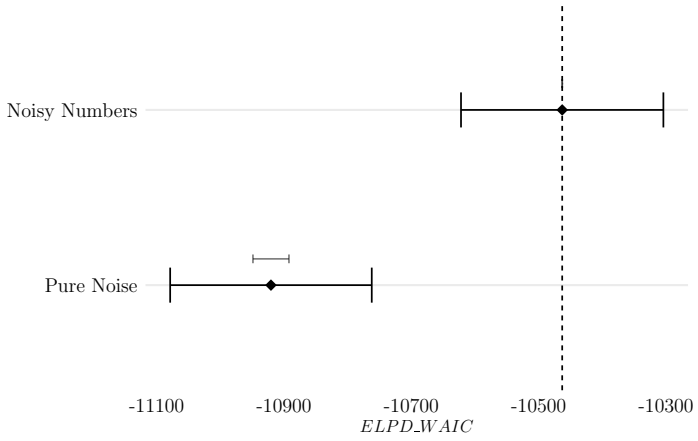


Table A.12 & Figure A.8: Model Comparison Number Comparison

A.3.2 Combined Estimation: Common and Separate Noise Terms

Common Noise Terms

	mean	median	sd	hdi 2.5%	hdi 97.5%	$\hat{R}$
Base Parameters:						
Altr. Preference β	0.297	0.297	0.014	0.269	0.324	1.01
Prior Std. Outcomes $\sigma^{\hat{r}}$	0.154	0.152	0.017	0.121	0.187	1.02
Prior Mean Outcomes $\mu^{\hat{r}}$	0.497	0.497	0.001	0.494	0.499	1.01
Prior Std. Preference $\sigma^{\hat{b}}$	0.273	0.263	0.061	0.173	0.398	1.01
Prior Mean Altr. Preference $\mu^{\hat{b}}$	0.512	0.511	0.027	0.462	0.567	1.01
Group Specific:						
Noise Baseline ν_B	0.277	0.276	0.024	0.232	0.326	1
Noise Treatment ν_T	0.338	0.337	0.031	0.281	0.401	1.01
Weight on Payments Baseline α_B	0.902	0.906	0.042	0.819	0.979	1.01
Weight on Payments Treatment α_T	0.863	0.867	0.056	0.753	0.966	1.01
Weight on Preferences Baseline γ_B	0.837	0.848	0.084	0.671	0.982	1.01
Weight on Preferences Treatment γ_T	0.779	0.788	0.106	0.575	0.971	1.01
Prior Treshold Baseline δ_B	0.958	0.96	0.08	0.79	1.112	1.01
Prior Treshold Treatment δ_T	0.946	0.945	0.107	0.73	1.152	1

Table A.13: Parameter Summary Combined Estimation. Jointly estimated parameter values of equation 7 and 9 based on 5000 posterior samples (+ 2000 warmup) per each of four chains. Parameters correspond to mean of log-normal hyper-distributions. Mean, median and sd refer to the mean, median and standard deviation of the posterior distribution draws. HDI 2.5% and HDI 97.5% indicate the borders of the 95% highest-density interval (HDI). $\hat{R}$ is a diagnostic of convergence of the Markov chains Markov chains ($\hat{R} = 1$ indicating convergence).

Separate Noise Terms

As stated, a combined dataset also allows to model two separate noise terms, i.e., $\nu_{\frac{self}{other}}$ and $\nu_{\frac{\beta}{1-\beta}}$ and identify the these domain-specific noise terms and the respective treatment effect. The combined and slightly adapted choice functions become:

$$Pr[self \succ other] = \Phi\left(\frac{\alpha \times \ln\left(\frac{self}{other}\right) - \gamma \times \ln\left(\frac{\beta}{1-\beta}\right) - \ln\left(\frac{\mu^{\hat{b}^{1-\gamma}}}{\mu^{\hat{r}^{1-\alpha}}}\right)}{\sqrt{\alpha^2 \times \nu^2_{\frac{self}{other}} + \gamma^2 \times \nu^2_{\frac{\beta}{1-\beta}}}}\right)$$
$$Pr[(A \succ B \times 1/2] = \Phi\left(\frac{\alpha \times \ln\left(\frac{A}{B}\right) - \ln\left(\frac{1}{2}\right) - \ln\left(\frac{1}{\mu^{\hat{r}^{1-\alpha}}}\right)}{\alpha' \times \nu_{\frac{self}{other}}}\right) \tag{A.1}$$

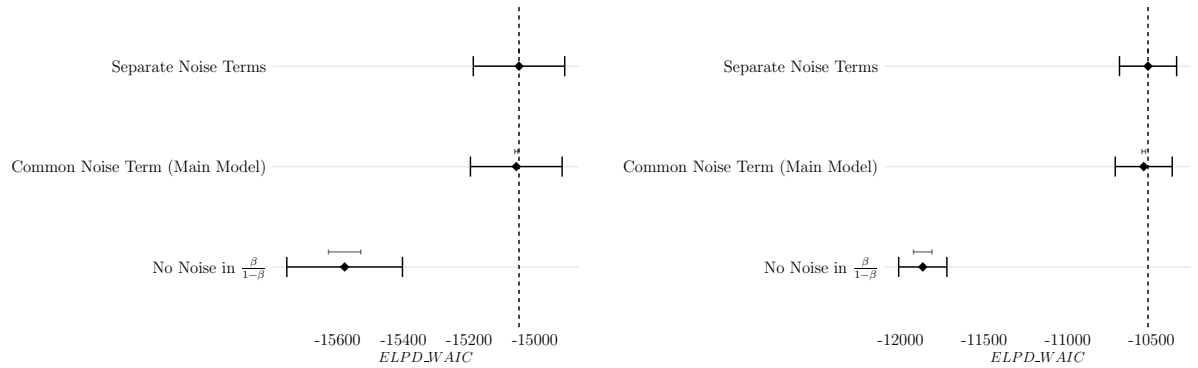
where $\alpha = \frac{\sigma^{\hat{r}^2}}{\sigma^{\hat{r}^2} + \nu^2_{\frac{self}{other}}}$ and $\gamma = \frac{\sigma^{\hat{b}^2}}{\sigma^{\hat{b}^2} + \nu^2_{\frac{\beta}{1-\beta}}}$. Again, the parameters relevant for number perception, $\nu_{\frac{self}{other}}$, $\sigma^{\hat{r}}$ and $\mu^{\hat{r}}$ are inferred from both number comparison and altruistic choices, whereas $\nu_{\frac{\beta}{1-\beta}}$, $\sigma^{\hat{b}}$, $\mu^{\hat{r}}$ and $\frac{\beta}{1-\beta}$ are separately estimated from only the altruistic choice data.

	mean	median	sd	hdi 2.5%	hdi 97.5%	$\hat{R}$
Base Parameters:						
Altr. Preference β	0.269	0.265	0.044	0.19	0.361	1.01
Prior Std. Outcomes $\sigma^{\hat{r}}$	0.527	0.515	0.091	0.368	0.708	1.06
Prior Mean Outcomes $\mu^{\hat{r}}$	0.466	0.468	0.017	0.429	0.496	1.02
Prior Std. Preference $\sigma^{\hat{b}}$	0.755	0.741	0.102	0.578	0.962	1.03
Prior Mean Altr. Preference $\mu^{\hat{b}}$	0.586	0.583	0.044	0.509	0.679	1.01
Group Specific:						
Noise Altruism Baseline $\nu_{\frac{\beta}{1-\beta},B}$	1.138	1.103	0.248	0.714	1.636	1
Noise Altruism Treatment $\nu_{\frac{\beta}{1-\beta},T}$	1.381	1.334	0.339	0.779	2.029	1
Noise Numbers Baseline $\nu_{\frac{self}{other},B}$	0.178	0.178	0.008	0.163	0.194	1
Noise Numbers Baseline $\nu_{\frac{self}{other},B}$	0.247	0.247	0.011	0.226	0.268	1
Weight on Payments Baseline α_B	0.89	0.894	0.035	0.824	0.951	1.06
Weight on Payments Treatment α_T	0.81	0.813	0.051	0.714	0.903	1.06
Weight on Preferences Baseline γ_B	0.32	0.311	0.104	0.136	0.53	1
Weight on Preferences Treatment γ_T	0.25	0.237	0.098	0.08	0.443	1.01
Prior Threshold Numbers Baseline $\delta_{\frac{self}{other},B}$	1.087	1.085	0.027	1.04	1.14	1.06
Prior Threshold Altruism Baseline $\delta_{\frac{\beta}{1-\beta},B}$	0.728	0.723	0.062	0.616	0.854	1.01
Prior Threshold Numbers Treatment $\delta_{\frac{self}{other},T}$	1.156	1.153	0.042	1.078	1.236	1.05
Prior Threshold Altruism Treatment $\delta_{\frac{\beta}{1-\beta},T}$	0.774	0.768	0.067	0.653	0.906	1.01

Table A.14: Parameter Summary Combined Estimation. Estimated parameter values of equation **??** based on 5000 posterior samples (+ 2000 warmup) per each of four chains. Parameters correspond to mean of log-normal hyper-distributions. Mean, median and sd refer to the mean, median and standard deviation of the posterior distribution draws. HDI 2.5% and HDI 97.5% indicate the borders of the 95% highest-density interval (HDI). $\hat{R}$ is a diagnostic of convergence of the Markov chains Markov chains ($\hat{R} = 1$ indicating convergence).

Model Comparisons

Model	Combined Choice Functions	$ELPD_{WAIC,A}$	$ELPD_{WAIC,NC}$
Separate Noise Terms	$Pr[self \succ other] = \Phi\left(\frac{\alpha\times\ln\left(\frac{self}{other}\right)-\gamma\times\ln\left(\frac{\beta}{1-\beta}\right)-\ln\left(\frac{\mu^{\frac{1-\gamma}{2}}}{\mu^{1-\alpha}}\right)}{\sqrt{\alpha^2\times\nu^2_{\frac{self}{other}}+\gamma^2\times\nu^2_{\frac{\beta}{1-\beta}}}}\right); Pr[(A \succ B \times 1/2] = \Phi\left(\frac{\alpha\times\ln\left(\frac{A}{B}\right)-\ln\left(\frac{1}{2}\right)-\ln\left(\frac{1}{\mu^{1-\alpha}}\right)}{\alpha\times\nu_{\frac{self}{other}}}\right)$	-15048.94	-10521.57
Common Noise Term (main Model)	$Pr[self \succ other] = \Phi\left(\frac{\alpha\times\ln\left(\frac{self}{other}\right)-\gamma\times\ln\left(\frac{\beta}{1-\beta}\right)-\ln\left(\frac{\mu^{\frac{1-\gamma}{2}}}{\mu^{1-\alpha}}\right)}{\nu\sqrt{\alpha^2+\gamma^2}}\right); Pr[(A \succ B \times 1/2] = \Phi\left(\frac{\alpha\times\ln\left(\frac{A}{B}\right)-\ln\left(\frac{1}{2}\right)-\ln\left(\frac{1}{\mu^{1-\alpha}}\right)}{\alpha\times\nu}\right)$	-15057.23	10547.44
No Noise in $\frac{\beta}{1-\beta}$	$Pr[self \succ other] = \Phi\left(\frac{\alpha\times\ln\left(\frac{self}{other}\right)-\ln\left(\frac{\beta}{1-\beta}\right)-\ln\left(\frac{1}{\mu^{1-\alpha}}\right)}{\nu\alpha}\right); Pr[(A \succ B \times 1/2] = \Phi\left(\frac{\alpha\times\ln\left(\frac{A}{B}\right)-\ln\left(\frac{1}{2}\right)-\ln\left(\frac{1}{\mu^{1-\alpha}}\right)}{\alpha\times\nu}\right)$	-15579.10	11867.78



(a) Altruism **(b)** Number Comparison

Table A.15 & Figure A.9: Model Comparison Combined Models. Altruism (a) and Number Comparison (b)

A.3.3 Robustness of Treatment Variation

Learning Effects and Fatigue First, I investigate the role of learning effects on altruistic and number comparison behavior. A straightforward way to do so is to augment the linear probability models of Tables A.2 and A.4 by a *Round* variable, which indicates in which of the 300 (200) rounds a decision was made. If the treatment effect is "learned", I expect a negative coefficient of the interaction effect between the treatment dummy and the round variable, i.e., a treatment effect that grows over time. The result of the corresponding linear probability model is depicted in Table A.16, where the first two columns refer to the altruism data and the last to the number comparison data. In the first two specifications, the coefficient of the interaction effect is indeed negative (-0.00006) and comparing round 0 to round 300 implies a 1.8 percentage point difference in selfish choices, which is sizable compared to the overall treatment effect. However, the coefficient is statistically insignificant ($p > 0.1$) in both specifications. In addition, if I take the results of column (1) at face value, already in round 0 the Treatment group decides 1.43 percentage points less often for $self$, which speaks against the responsibility of learning effects for the treatment difference. I arrive at a similar conclusion, albeit with different evidence, for the number comparison data: In columns (3) and (4), I include the mentioned interaction effect. I observe a statistically significant *positive* coefficient of the interaction effect of (0.00013), which implies an *increase* in 2.6 percentage points to choose A between round 0 and round 200. Instead of growing over time, this implies that the treatment effect shrinks. Supporting this argument is that in round 0, the treatment effect is sizable and statistically significant and the Treatment group decides 4.916, respectively 5.84 percentage points less for A. The four columns thus do not provide evidence for a learned treatment effect and instead, point towards some attenuation over time in the number comparison data.

	(1)	(2)	(3)	(4)
Treatment Group	-0.01430**	-0.01430	-0.04916***	-0.05084***
	(0.00603)	(0.02342)	(0.00362)	(0.00690)
Ratio $\frac{self}{other}$ / $\frac{A}{B}$	0.86707***	0.86707***	-0.87519***	1.32264***
	(0.00372)	(0.01323)	(0.00184)	(0.00736)
Treatment Group * Round No.	-0.00006	-0.00006	0.00012***	0.00013***
	(0.00004)	(0.00006)	(0.00003)	(0.00005)
Intercept	0.05165***	0.05165***	1.78986***	-0.18654***
	(0.00459)	(0.01862)	(0.00390)	(0.00444)
Num.Obs.	72000	72000	60000	60000
Data	Altruism	Altruism	Number Comp.	Number Comp.
Clustered Standard Errors	No	Yes	No	Yes
Random Effects	No	Yes	No	Yes
Unique Obs	300	300	300	300
R2	0.430	0.517	0.791	0.679

Table A.16: Main Treatment Effect and Learning Regression. Linear Probability Model. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky and Tipton, 2018)) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Two other facts that are insightful for learning over time come from the decision in the *very first* illustrative example as well as the 12 consecutive practice trials. At the very beginning and as part of explaining the study, participants had to make a non-consequential decision whether to take 2.31 € (= 1.72 € + 0.59 €) for themselves or give

4.66 € (= 1.14 € + 3.52 €) to another person. In this decision, there is no treatment difference as $\bar{self}_T = 0.393$, $\bar{self}_B = 0.367$ ($p = 0.6356$). However, in the 12 practice trials[32] the Treatment group decides significantly more often for the other person with an average of $\bar{self}_T = 0.364$, $\bar{self}_B = 0.414$ ($p < 0.01$). Thus, during the practice trials, the Treatment group is much more pro-social. While I acknowledge limits for drawing conclusions from this data – given it is non-incentivized, only for practice purposes and contains only 10 decisions – this behavior would be consistent with the following interpretation: Participants quickly understand how the task works, i.e., "less-for-me" vs "more-for-other". This, in turn, could translate into the intuition that $\mu^{\hat{r}} < 1$ and thus a higher pro-sociality of the Treatment group (throughout the experiment). All in all, I thus interpret the presented data as evidence that the treatment effect is not learned over the repeated trials of the experiment but that participants formed a quick intuition about the rules of the task.

The treatment variation could also introduce differences in *fatigue levels*, which are responsible for the difference in choices between both groups. If differences in fatigue are not already present in the very first choices, the above analysis already provides some evidence against this argument. In addition, I can test both for group differences in (i) revealed (effects of) fatigue and (ii) subjectively reported fatigue levels. Regarding the first, often-discussed consequences of fatigue are more errors in choices and, higher levels *choice inconsistency*, an argument especially relevant for survey design (see e.g., Bech et al. (2011); Schwappach and Strasmann (2006); Özdemir et al. (2010)). Our data offers a unique way of analyzing the determinants of choice inconsistency: Recall that each trial of the altruism and number comparison task was repeated five times (which is one game), yet the order of trials was randomly determined. This implies that some participants encountered the fifth iteration of a given game earlier in the experiment compared to other participants, which induces exogenous differences in the completion rounds of a given trial group. If fatigue (differences) increase throughout the experiment, *later completions* should be associated with a higher choice inconsistency.

	(1)	(2)	(3)	(4)
Trial Final Round No.	-0.00005	-0.00006	-0.00005	-0.00005
	(0.00007)	(0.00006)	(0.00008)	(0.00007)
Trial Final Round No. * Treatment Group	0.00000	0.00003	0.00001	0.00006
	(0.00009)	(0.00009)	(0.00011)	(0.00010)
Intercept	0.11693***	0.04352	0.10210***	0.03006
	(0.03063)	(0.03111)	(0.03079)	(0.02785)
Data	Altruism	Altruism	Number Comp.	Number Comp.
Trial Group Fixed Effects	No	Yes	No	Yes
Participant Fixed Effects	Yes	Yes	Yes	Yes
Num.Obs.	14400	14400	12000	12000
R2	0.131	0.206	0.066	0.309

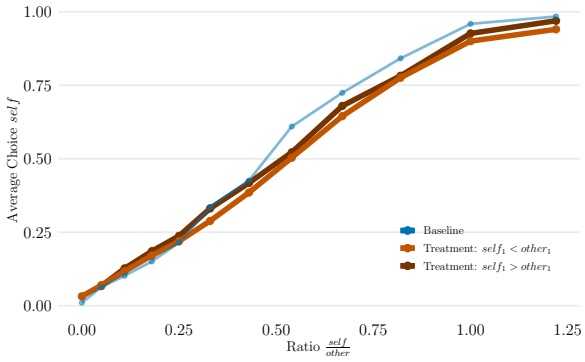
Table A.17: Inconsistency Regression on Trial Level. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky and Tipton, 2018)) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table A.17 performs a linear regression on a dataset with 300 (subject) × 48 (trial

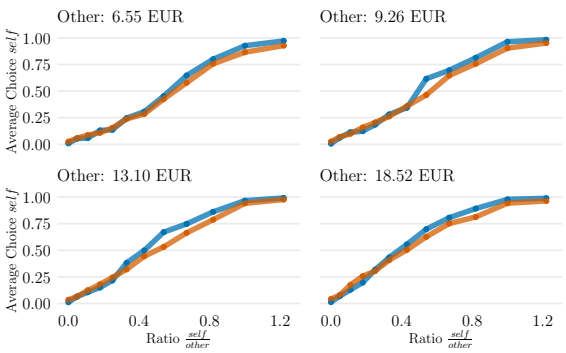
[32] Recall that in the practice trials, I fixed *other* = 10.00 € and varied *self* ∈ 0, 0.52, 1.11, 1.76, 2.50, 3.33, 4.28, 5.38, 6.66, 8.18, 10.00, 12.22 €, i.e., these trials in principle already allow to infer something about β.

groups, see Figure A.4) = 14,400 respectively $200 \times 40 = 12{,}000$ observations based on the Altruism and Number Comparison data. Each row of this dataset contains the standard deviation of a given trial group by a given participant alongside its completion round, i.e., in which round a participant encountered the fifth and last iteration of that trial group. Crucially, this dataset allows for the inclusion both of participant and trial group fixed effects. Every specification in Table A.17, regardless of the data source, does neither provide evidence for a growing inconsistency, nor a growing difference in inconsistency between both groups. Thus, participants' choices do not get more inconsistent over time nor does the choice inconsistency develop differently between the treatment and control group. In addition to implied fatigue effects, I collected self-reported measures of fatigue levels using visual analog scales (Radbruch et al., 2003). I asked participants both about their current level of fatigue as well as the average during the past 24 hours on a scale of 0-10 using a slider (see Figure A.18). The Treatment group indeed does report slightly higher levels of current fatigue ($\bar{fatigue}_T = 4.775, \bar{fatigue}_B = 4.310$, $p = 0.1005$). However, it is not obvious that higher levels of self-reported fatigue necessarily translate into different choices. I will pick up this point in more detail in Section A.3.4 during the estimation of heterogeneous treatment effects.

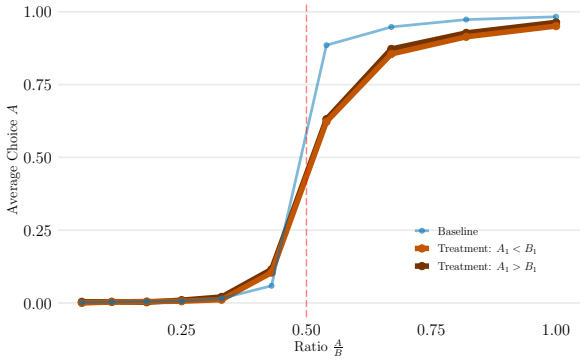
Mechanical Difference in Choices and Increase in Inconsistency Alternative to the above-mentioned points, an alternative explanation for the treatment effect could be a purely "mechanical" one: If participants only focus on the *first components* of the sums in the Treatment group and simply pick the larger they would behave both more variable due to the random placement of the position of the components and behave less selfish if the first component of the *other* variable ($other_1$) is larger more often. If this argument holds, I should observe a higher level of selfishness (compared to Baseline) if $self_1 > other_1$ and a lower level once $self_1 < other_1$.



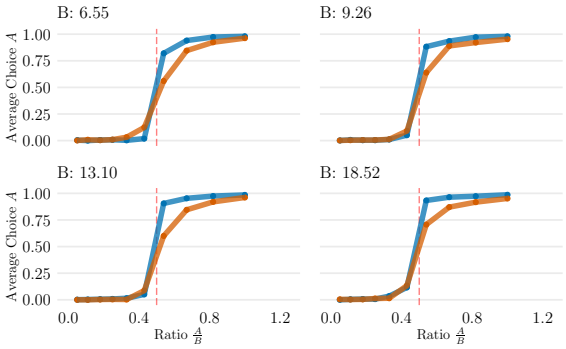
(a) Math components: Altruism



(b) Stake sizes: Altruism



(c) Math components: Number Comparison



(d) Stake sizes: Number comparison

Figure A.10: Components of To-Be-Calculated Sums and Stake Sizes. Panels (a) and (c) plot the average choice for $self$, respectively A as a function of $\frac{self}{other}$, $\frac{A}{B}$, for Baseline and Treatment whereas the latter is divided into cases where the first math component $self_1 > other_1$ and $self_1 < other_1$, with $self = self_1 + self_2$ and $other = other_1 + other_2$. Panels (b) and (d) plot average choices separately for the different base values of $other$ and B.

Panel (a) of Figure A.10 plots the average choice for $self$ separately depending on the numerical configuration of the math components, i.e., if $self_1 > other_1$ or $self_1 < other_1$. I observe comparable differences to the Baseline group within the Treatment trials regardless of the relationship between $self_1$ and $other_1$. While indeed $other_1 > self_1$ (58.95 %) occurs more frequently than $other_1 > self_1$ (41.05 %), the fact that participants still behave *more* pro-social compared to Baseline in both groups of trials speaks against a purely "mechanical" increase in pro-sociality. In the number comparison task (panel c), I observe virtually no difference in behavior between trials where $A_1 > B_1$ and $A_1 < B_1$.

Another explanation could be that the treatment variation perhaps only works for smaller values of $self, other$ where the sums will generally contain smaller values. For example, one could expect a stronger treatment effect in trials such as $self = 3.52$ (= 1.61 + 1.91) vs $other = 6.55$(=4.86 + 1.69) compared to $self = 7.05$ (= 4.42 + 2.63) vs $other = 13.1$(=10.32 + 2.78) as the components of the sums in the former set of trials are simply smaller (while the ratio between $self$ and $other$ remains the same). Thus, the overall treatment effect could be driven by the impact on trials with generally smaller math components which are more likely to be disregarded by participants. Panel (b) of Figure A.10 plots the average choice for $self$ separately for the four different levels of stakes in the trials. Even though there is some difference in behavior between the different stake groups, the *treatment effect* is very similar across different stakes. The same is true for the number comparison task (panel (d)), where I observe similar treatment effects regardless of the value of the number B. Table A.18 performs a regression analysis akin

to the linear probability models in Tables A.2 and A.4 and shows that the treatment effect does not systematically depend on the general stakes of the trial (or the value of *other* and B), and also shows that larger stakes are associated with more choices for $self$ and A.

	(1)	(2)
Game Group (1-4)	0.03749***	0.00875***
	(0.00233)	(0.00249)
Game Group (1-4) * Treatment Group	-0.00009	-0.00533
	(0.00330)	(0.00352)
Treatment Group	-0.02222**	-0.02400**
	(0.00903)	(0.00965)
Intercept	0.35806***	0.36627***
	(0.00638)	(0.00682)
Num.Obs.	72000	60000
Data	Altruism	Number Comp.
Clustered Standard Errors	Yes	Yes
Unique Obs	300	300
R2	0.008	0.002

Table A.18: Stakes-Size and Choice Regression. Linear Probability Model. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky and Tipton, 2018)) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

A.3.4 Heterogeneous Treatment Effects

To further investigate the nature of the treatment effect, I analyze heterogeneous treatment effects. To do so, I leverage recent developments in the causal machine learning literature and employ a Causal Forest for estimating heterogeneous treatment effects (Wager and Athey, 2018). Causal Forests adapt the logic of tree-based models identifying a split at a given level of a covariate to minimize a loss criterion to the estimation of treatment effects and search for splits that maximize heterogeneity in the estimated conditional average treatment effect. Importantly, these methods are "honest", i.e., use a different set of data points to propose and evaluate the splits. I use the CausalForest class implemented in the `econml` package (Battocchi et al., 2019).[33] The main advantage over classical techniques (i.e., interaction term in OLS) is that they – constructed using cross-fitting techniques – are less prone to overfitting and able to pick up other functional forms beyond linear or explicitly pre-specified ones.

[33]Note that I can define the treatment propensity model – what usually needs to be estimated from the data – as a fair coin flip given the exogenous treatment assignment in the experiment.

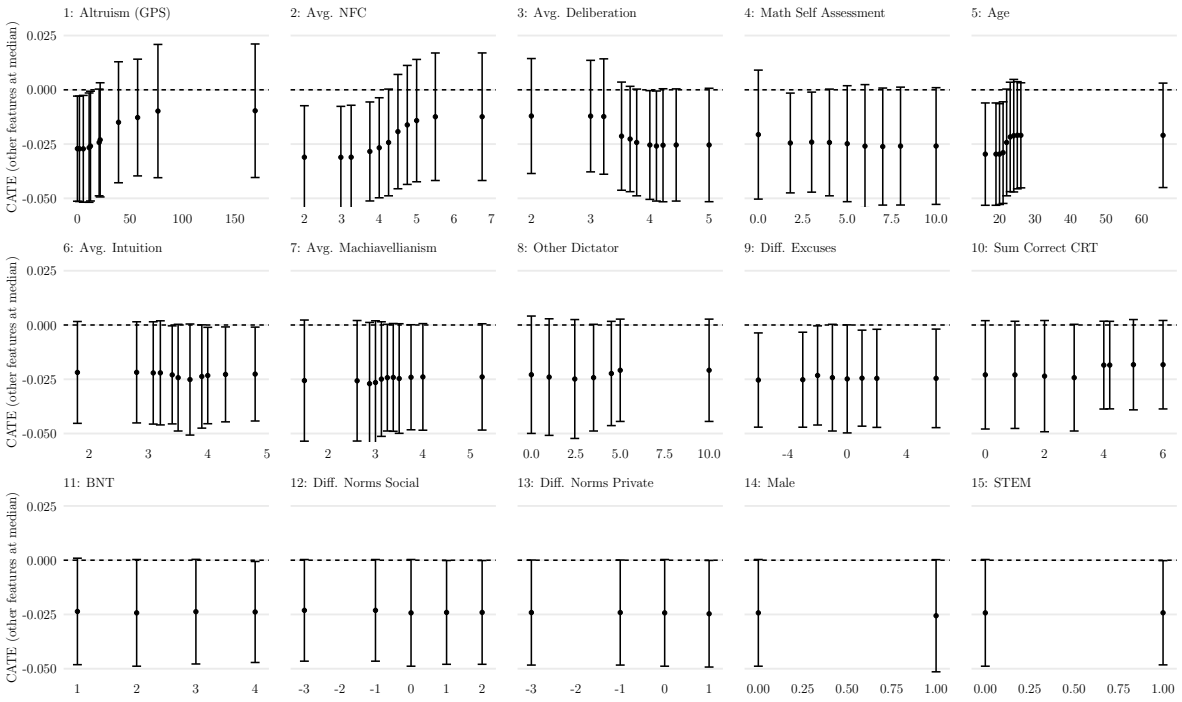


Figure A.11: Heterogeneous Treatment Effects. This plot shows the estimated CATE values for each personal characteristic (sorted by their importance), estimated at each decile of the feature distribution, 95 % confidence intervals.

Figure A.11 shows the estimated CATE values for the different personal characteristics (while holding the remaining characteristics at their median value), sorted in descending order by their importance for the estimated CATE values. Key take-away is that (i) the variation in most personal characteristics do not contribute meaningfully to the CATE estimates and only participants high on the self-reported GPS Altruism score seem to be slightly less impacted by the treatment compared to lower-scoring individuals. Similarly, participants high on the Need for Cognition scale are less impacted by the treatment, but the confidence intervals are relatively large in both cases. For the remaining features, most variation corresponds to the average treatment effect (ATE). This is true for the remaining "cognitive" measures, such as the CRT or BNT performance, which do not indicate systematic treatment heterogeneity. Importantly, further is that personal characteristics that could be related to a tendency to "exploit" potential side-effects of the chosen treatment variation, i.e., how strongly their self-reported negative emotions after selfish behavior react to the availability of excuses, the difference between private and social norms in the treatment and the average score on the Machiavellianism scale are not major sources of heterogeneous treatment effects. This is further evidence that the treatment effect – in addition to leading to *less* selfish choices – did not invoke motivated "second-order" behavior. Overall, the estimation of heterogeneous treatment effects leads to the conclusion that the treatment effect does not operate systematically differently for participants depending on their characteristics and that our dataset is not large enough to detect minuscule differences in treatment heterogeneity.

A.3.5 Correlation between Metacognition, RT and Choices

Table A.19 in the appendix shows pairwise rank correlation coefficients between the various metacognitive measures and average choices for $self$, standard deviation (on a game level), and time spent in altruism choices as well as the average correct choices, standard deviation and time spent in the number comparison task. The main finding is that there is astronger association between choices, inconsistencies and RT with metacognitive measures in the number comparison domain compared to the altruism choices.

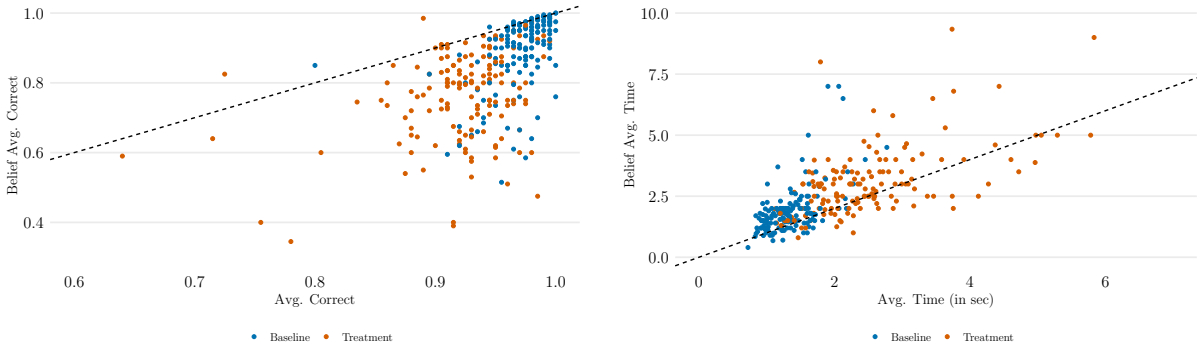
	Altruism (Avg.)			Number Comparison (Avg.)		
	Choice *Self*	Std. Deviation	RT	Choice Correct	Std. Deviation	RT
Altruism:						
Negative Confidence	0.088	0.317***	0.080	-0.127*	0.109	0.048
Avg. Attention	-0.215***	-0.045	0.115*	0.094	-0.078	-0.066
Precision	-0.195***	-0.114	0.126*	0.006	-0.005	-0.040
Number Comparison:						
|Δ Belief Correct|	-0.057	0.257***	0.198***	-0.365***	0.392***	0.433***
Belief Correct Confidence	0.043	-0.161**	-0.203***	0.309***	-0.285***	-0.365***
|Δ Belief Time Spent|	0.074	0.107	0.125*	-0.077	0.040	0.209***
Belief Time Spent Confidence	0.033	0.003	-0.121*	0.025	-0.031	-0.151**
Precision	0.020	-0.017	-0.113	0.254***	-0.256***	-0.150*
Avg. Attention	0.008	-0.131*	-0.093	0.223***	-0.231***	-0.058

Table A.19: Correlation Metacognition in Altruism and Number Comparison. p-values from pairwise rank-correlation tests ($n = 300$). * p < 0.1, ** p < 0.05, *** p < 0.01.

Starting with the upper-left quarter of the table, I observe that subjects who decide less often for $self$ also report higher levels of attention ($\rho = -0.215$) and precision ($\rho = -0.195$), yet there is no apparent correlation of altruistic choices with the confidence measure. Proceeding to the second column, I do observe a positive correlation between the average standard deviation and confidence ($\rho = 0.317$): Participants who report a lower level of confidence are more inconsistent in their altruistic behavior (but not more or less pro-social on average). The average time spent on making altruistic choices does not meaningfully correlate with any of the metacognitive measures. Notably, these correlations also indicate that the metacognitive measures do not replicate the treatment effect: One could expect that "less metacognitive" participants also tend to decide more often for *other* given the direction of the treatment effect towards fewer choices for $self$, but this appears not the be the case. Together with the previous fact that there are no treatment differences in the altruism metacognitive measures, this implies that the mechanism through which the treatment effect operates is likely not via impacts on (conscious) metacognition.

The behavioral data from the altruism domain also correlate to some extent with measures of metacognition *across domain* as shown in the lower-left quarter: Participants with a larger |Δ Belief Correct|, i.e., whose beliefs deviate more from their true performance and lower confidence in their belief statements are more inconsistent in their altruism decisions ($\rho = 0.257$; $\rho = -0.161$) and take longer to choose between $self$ and *other* ($\rho = 0.198$; $\rho = -0.203$). This reiterates the argument in Section 4.7.1 that, altruism and number comparisons, at least to some extent, are driven by similar processes if elicited in comparable settings.

Turning to the upper-right quarter of Table A.19, I observe no apparent correlations between the altruism metacognitive measures and number comparison behavior. Within-domain, this is different: As shown in the lower-right quarter, the fewer correct choices a participant makes, the more their beliefs deviate from their true performance ($\rho = -0.365$), the less confident they are in their belief estimates ($\rho = 0.309$), the lower the self-reported precision ($\rho = 0.254$) and attention ($\rho = 0.223$). Inconsistency in the number comparison also correlates with belief deviations ($\rho = 0.392$), their confidence ($\rho = -0.285$) as well as self-reported precision ($\rho = -0.256$) and attention ($\rho = -0.231$). And finally, also the time spent is larger the more a participant deviates in their belief statements ($\rho = 0.433$), lower the higher the confidence ($\rho = -0.365$), higher the more a participant deviates in their belief statements of decision time ($\rho = 0.209$) and higher the lower the confidence in these statements ($\rho = -0.151$).

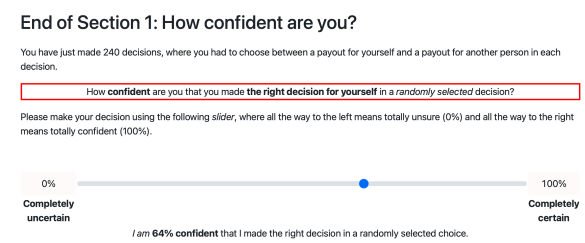


(a) Beliefs (Avg.) Correctly Solved Tasks

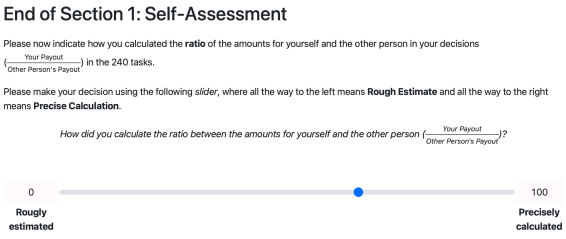
(b) Beliefs (Avg.) Time Spent

Figure A.12: Beliefs Number Comparison

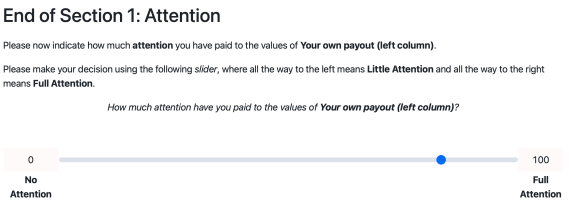
A.4 Experimental Screenshots



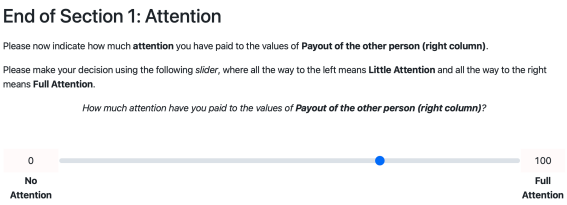
(a) Confidence



(b) Precision



(c) Attention *self*



(d) Attention *other*

Figure A.13: Screenshots: Metacognition Altruistic Choice.

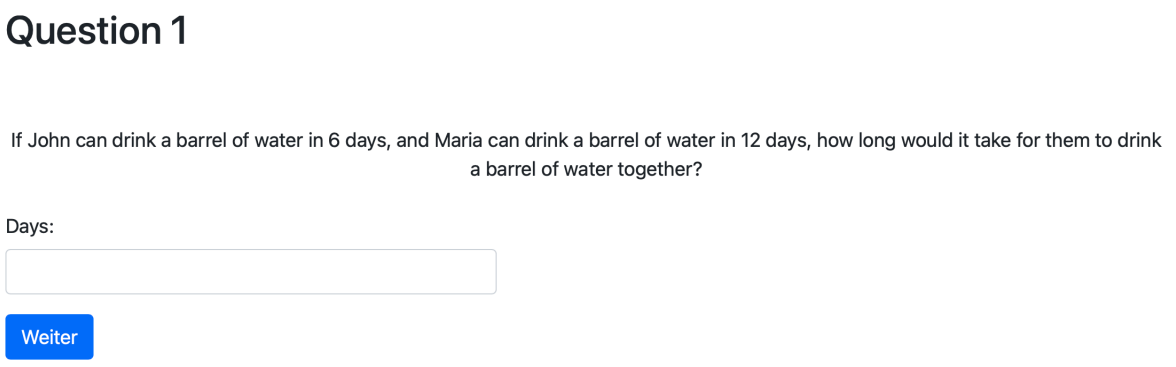
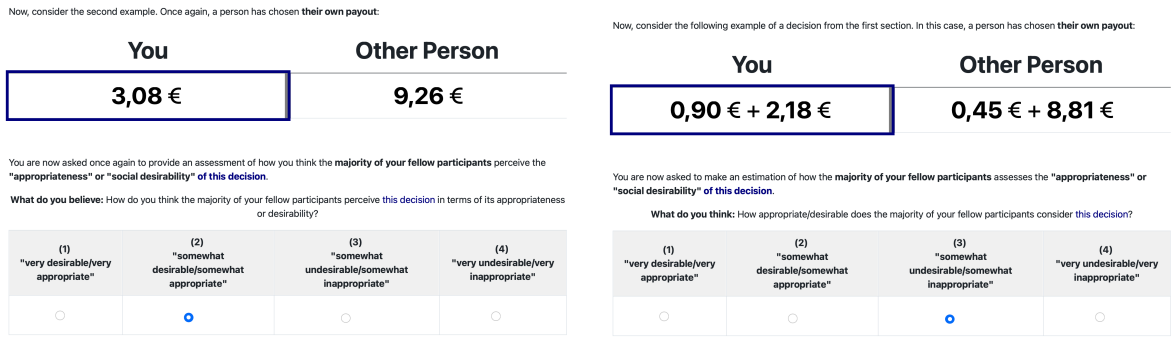


Figure A.14: Screenshot: CRT4 Question.

(a) No Excuses

(b) Excuses (reverse formulated)

Figure A.15: [Screenshots: Excuse-Taking Questions. Survey Questions inspired by Lepper (2024). **(a)** No excuses **(b)** Excuses. Order in which questions appear is randomized.



(a) Baseline (b) Treatment

Figure A.16: Screenshots: Social Norms. **(a)** Baseline **(b)** Treatment. Both variants are shown to all participants, order in which questions appear is randomized

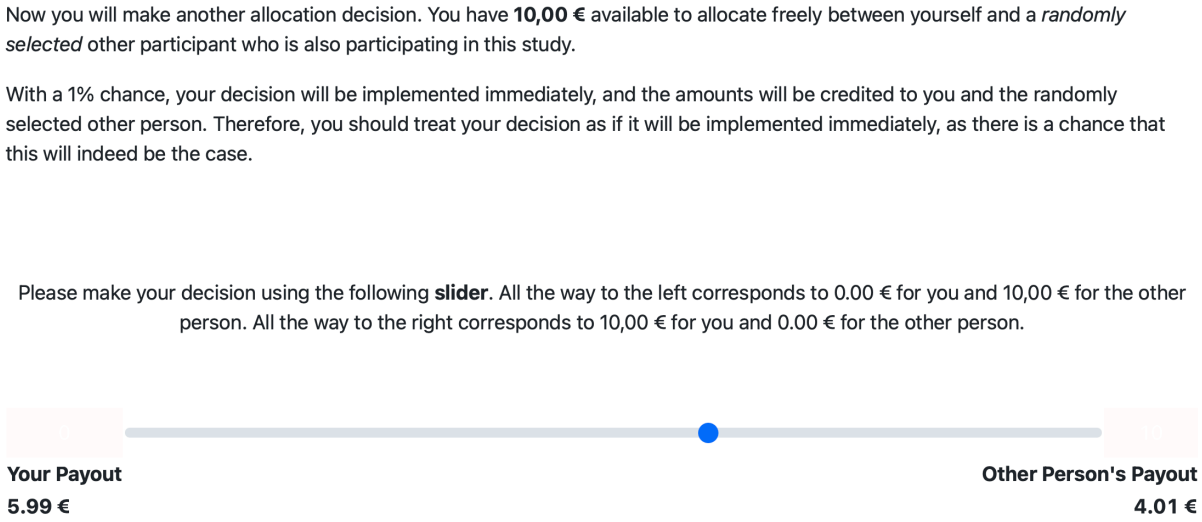


Figure A.17: Screenshot: Dictator Game.

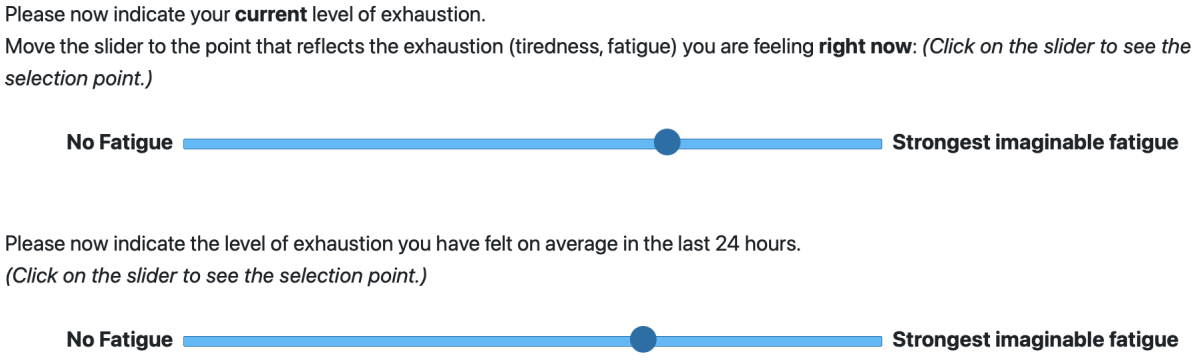


Figure A.18: Screenshot: Fatigue Visual Analog Scales.