Gutenberg School of Management and Economics

& Research Unit “Interdisciplinary Public Policy”

Discussion Paper Series

Cognitive Noise and Altruistic Preferences

Niklas M. Witzig

September 2024

Discussion paper number 2415

Johannes Gutenberg University Mainz
Gutenberg School of Management and Economics
Jakob-Welder-Weg 9
55128 Mainz
Germany
https://wiwi.uni-mainz.de/

Contact details

Niklas M. Witzig
Chair of Public and Behavioral Economics
Johannes Gutenberg University Mainz
55128 Mainz
Germany
niklas.witzig@uni-mainz.de

Cognitive Noise and Altruistic Preferences*

Niklas M. Witzig**[a]

[a] *Johannes Gutenberg University of Mainz*

This version: December, 2024

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 math sums replace straightforward 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 the perception of numerical magnitudes drives treatment differences. The probabilistic model supports 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 the expression of altruistic preferences – and potentially social preferences more generally – is affected by the cognitive difficulty of their implementation.

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

*I thank Alexander Dzionara, Markus Eyting, Ben Grodeck, Katharina Hartinger, Florian Hett, Marc Kaufmann, Sebastian Olschewski, Daniel Schunk, Ferdinand Vieider 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.

**Corresponding author (niklas.witzig@uni-mainz.de).

1 Introduction

Theories of social and other-regarding preferences characterize a crucial advancement in economics and help to explain the results of various laboratory and field outcomes irreconcilable with traditional assumptions of pure self-interest (Andreoni & Miller, 2002; Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Fehr & Schmidt, 1999; Fisman et al., 2007; Levine, 1998), with reviews in Cooper and Kagel (2016), Fehr and Charness (2023), and Fehr and Schmidt (2006). Quantifying the underlying motivations of prosocial behavior, a growing body of research estimates population- and individual-level parameters of different social preference frameworks (e.g., Bellemare et al., 2011; Bruhin et al., 2019; Carpenter & Robbett, 2024; Fisman et al., 2007; Klockmann et al., 2022 with a meta-analysis of inequality-aversion estimates available in Nunnari & Pozzi, 2024).

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 utility-maximizing 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 contrasts 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 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 solutions as an immediate consequence of such noisy perceptions (Frydman & Jin, 2022; Khaw et al., 2021; Vieider, 2024b; Woodford, 2012, 2020). 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 & Graeber, 2023; Enke et al., 2023; Oprea, 2024).

It is only natural to assume that social preferences are affected by cognitive imprecisions and the complexity of their implementation. 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 will 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

1 Oprea (2024, p. 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."

in previous work – are a candidate, but not-yet considered driver of choices related to social preferences.

First indicative evidence that noisy cognition or complexity-related processes can guide prosocial choice is now starting to amass: For example, Enke et al. (2024) consider the dictator game as one example of how cognitive uncertainty moderates reactions to changes in objective problem features across over 30 experiments. Similarly, Bao and Pei (2024) find that higher cognitive uncertainty is associated with higher contributions in the public goods game. Beyond that, empirical evidence for noisy cognition (or complexity-related effects more generally) on social preferences is still lacking, however, particularly, in domains with no clear "default-action".

In this paper, I investigate altruistic choices – a simple form of social preference-related decisions – through the lens of a cognitively noisy decision-maker. Based on Vieider (2024b), I develop a model of altruistic choice and show how an increase in cognitive noise can directionally affect altruistic choices. The core intuition is that higher cognitive noise – either in perceiving monetary payments or altruistic preferences – 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 reacts less strongly to changes in underlying problem features and, depending on the mental default, also chooses systematically differently.

To test these implications, I implement a laboratory experiment that consists of 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., decide between $\text{self}_1 + \text{self}_2$ (= self) and $\text{other}_1 + \text{other}_2$ (= other). Encasing the stakes in to-be-calculated sums increases the cognitive difficulty of perceiving the monetary payments and deciding on 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 aims to mirror 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 indicates a considerable degree of uncertainty around

the mechanism of the treatment effect, suggesting that additional data beyond altruistic decisions is necessary to make a precise statement about the origin of the treatment effect. Herein lies the main contribution of 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 $\widehat{\text{self}} < \widehat{\text{other}}$ and $\widehat{\text{A}} < \widehat{\text{B}}$, is a candidate driver for the treatment effect in both tasks. This conclusion is supported by probabilistic models based on both the number comparison data and on a *combined* dataset of behavior in both tasks, indicating a high probability of such an "intermediate" prior belief of numerical magnitudes.

In 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 measuring altruistic behavior more generally. Further, individual parameter estimates (based on the hierarchical models) of cognitive noise correlate with performance on the Cognitive Reflection Test and Berlin Numeracy Task, showing that more cognitively able persons are also less cognitively noisy, providing support for the cognitive motivation of the general framework. More exploratory analyses show how measures of meta-cognition (e.g., self-reported confidence and attention) and response times – both key informants 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 the recent work of the cognitive turn in behavioral economics. Most of this work so far focuses on the domain of risk, ambiguity, belief updating and intertemporal choice (Enke et al., 2023; Frydman & Jin, 2022; Khaw et al., 2021; Vieider, 2024a; Vieider, 2023; Woodford, 2020). 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 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. (2023), provide an easy-

to-implement method of increasing uncertainty in the perception of objective problem features that also have proven to be suitable in a more extensive repeated-trials experiment. Employing exogenous manipulations further speaks to a broader ongoing discussion in this literature: Enke (2024, p. 57) outlines how it is often unclear which assumptions to make about the (locations of the) prior distributions in the Bayesian models. Here, I show that typical ignorance assumptions are not necessarily valid (see also p. 33 Oprea & Vieider, 2024) and demonstrate how a combination of experimental variations increasing noise, "mirror" tasks isolating parts of the decision-making process and probabilistic modeling allow inferring the parameters of prior distribution and likelihood in the Bayesian models.

This paper also relates to the literature on structural estimations of social preference parameters (Bellemare et al., 2011; Bruhin et al., 2019; Carpenter & Robbett, 2024; Echeverry et al., 2023; Fisman et al., 2007; Klockmann et al., 2022; Nunnari & Pozzi, 2024). Here, I show how altruistic behavior (and thereby "revealed altruistic preferences") can be affected by an increase 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; Carpenter & Robbett, 2024; Van Leeuwen & Alger, 2024) or using estimated social preferences are used to predict or related to real-world outcomes (e.g., as in Graf et al., 2013) potentially suffers from biases. Furthermore, this paper makes an additional 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 modeling more generally.

Lastly, this paper relates to an interdisciplinary literature on dual-process models of cognition, 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 altruistic choices in the Treatment group 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 altruistic choices independent of altruistic preferences. 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 altruistic choices 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 5 discusses the main results of the paper, outlining potential avenues for future research while Section 6 briefly concludes, highlighting the limitations of the current paper.

2 Theoretical Framework

The theoretical model modifies models of noisy Bayesian cognition by Vieider (2024b) and Khaw et al. (2021) and applies them to choices involving 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 choose self if

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

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 & Stark, 1988; Levine, 1998). While the value of β can, in principle, be any real number, a sensible restriction is to expect $\beta \in (0, 0.5)$, i.e., that the DM places 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 to social preferences, such as (dis-)advantageous inequality aversion (Fehr & Schmidt, 1999), reciprocity concerns (Bellemare et al., 2011; Falk & Fischbacher, 2006), or social norms (Carpenter & Robbett, 2024) and also does

not distinguish between (non-)warm-glow giving (Andreoni, 1989). Instead, this rule 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 & Miller, 2002; Charness & 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:

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

which states that the DM assesses whether the (log) ratio of monetary payments, $\ln\frac{\text{self}}{\text{other}}$, 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. For example 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 discussed more extensively below) which in turn will be relevant to many choice rules featuring a comparison of monetary payments.

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 – $\frac{\text{self}}{\text{other}}$ 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 (Dehaene, 2011; Nieder & Dehaene, 2009).[4] This

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 can be linked to the idea of a "Bayesian Brain" (Doya et al., 2006) from neuroscience, i.e., that the brain optimally combines uncertain sensory evidence with prior knowledge.

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 & Nieder, 2007).

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, p. 2219). These observations suggest 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 conjecture, studies aiming to map the "mental number line" also find evidence for a non-linear compressed mental representation of Arabic numerals (Longo & Lourenco, 2007). Furthermore, Prat-Carrabin and Woodford (2022) show that the relationship between discriminability and bias – a core law of human perception (Wei & 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 $\frac{\text{self}}{\text{other}}$ and $\frac{\beta}{1-\beta}$, assume that the DM obtains mental signals about the true values from a distribution of possible representations:

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

where $s_{\frac{\text{self}}{\text{other}}}$ and $s_{\frac{\beta}{1-\beta}}$ are the mental signals. Importantly, I do not assume that the mental signals share a common variance, but instead that $\nu_{\frac{\text{self}}{\text{other}}}$ characterizes "noise in monetary payments" and $\nu_{\frac{\beta}{1-\beta}}$ characterizes "noise in altruistic preferences". Cognitive noise, as understood here, thus consists of two different sources of noise in perceiving problem features. This departs from previous work, where a common noise variance is a typical assumption (Vieider, 2024b). The main argument for separately modeling the noise terms in the present setting is that $s_{\frac{\text{self}}{\text{other}}}$ and $s_{\frac{\beta}{1-\beta}}$ do not both refer to an explicitly stated numerical quantity, like e.g., a lottery payoff and probability, but to monetary payments and a *subjective* preference threshold. Assuming that the same (cognitive) process underlies the perception of both features is thus less justified in the present setting.

What remains common to both noise terms is the overall log-normal noise structure. With both means as logarithms, the noise terms become 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).

In addition to the mental signals, the Bayesian DM has *prior beliefs* about both the ratio of monetary payments and the preference threshold.

$$\ln \frac{\text{self}}{\text{other}} \sim \mathcal{N}\left(\ln \mu_{\hat{r}}, \sigma_{\hat{r}}^2\right), \ln \frac{\beta}{1-\beta} \sim \mathcal{N}\left(\ln \mu_{\hat{b}}, \sigma_{\hat{b}}^2\right) \tag{4}$$

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

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

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

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

with the following expected values:

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

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

where $\alpha = \frac{\sigma_{\hat{r}}^2}{\sigma_{\hat{r}}^2 + \nu^2_{\frac{\text{self}}{\text{other}}}}$ and $\gamma = \frac{\sigma_{\hat{b}}^2}{\sigma_{\hat{b}}^2 + \nu^2_{\frac{\beta}{1-\beta}}}$, the Bayesian evidence weights. 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 signals of the) true values of $\frac{\text{self}}{\text{other}}$ and $\frac{\beta}{1-\beta}$.

So far, this setup follows a common structure in the noisy cognition literature. Incorporating a prior belief for monetary payments – with the values $\frac{\text{self}}{\text{other}}$ varying from trial to trial in a typical experiment – leads to a regularization in the posterior belief as

shown above and typically assumed. However, altruistic preferences are again conceptually somewhat different: First, $\frac{\beta}{1-\beta}$ is usually assumed to be a fixed quantity for a given DM, which in turn makes it difficult (although not impossible) to distinguish between true and prior preferences. From an empirical perspective, this translates into identification challenges and risks of overly parameterizing the later choice model. Therefore, I adjust this common setup and abstract from any additional influence of the prior belief over preferences on choices. Therefore, throughout, I assume that the prior belief is maximally uninformative in terms of inference over true preferences. Importantly, noise in the mental signals $s_{\frac{\beta}{1-\beta}}$ still impacts choices (discussed in more detail below) and this assumption does not imply that perception of preferences in noiseless, only that noise in perceiving preferences does not lead to a bias towards some mental default preference. In terms of the theoretical framework, by setting $\sigma_{\hat{b}} \to \infty$, this implies that $\lim_{\sigma_{\hat{b}} \to \infty} \gamma = 1$ and, in turn $E\left[\ln\left(\frac{\beta}{1-\beta}\right) \mid s_{\frac{\beta}{1-\beta}}\right] = s_{\frac{\beta}{1-\beta}}$. Therefore, the prior belief regarding altruistic preferences does not affect the posterior distribution (or expectation); only the monetary payment prior fulfills the typical regularizing role.

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

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

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

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

where $\delta = \frac{1}{\mu_{\hat{r}}^{1-\alpha}}$. The DM decides for self if the difference between the weighted signal of monetary payments and the signal of their altruistic preference is larger than a prior-induced threshold. To arrive at a probabilistic choice rule, subtract the z-score of the random variable $\alpha \times s_{\frac{\text{self}}{\text{other}}} - s_{\frac{\beta}{1-\beta}} \sim \mathcal{N}(\alpha \times \ln \frac{\text{self}}{\text{other}} - \ln \frac{\beta}{1-\beta}, \nu^2_{\frac{\text{self}}{\text{other}}} \alpha^2 + \nu^2_{\frac{\beta}{1-\beta}})$ from the equivalent z-score of equation 6, which results in the following Probit equation (see Vieider (2024b) for the original proof):

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

Choosing self is thus the outcome of a *probabilistic* process in which both noise in monetary payments – and a potentially invoked bias of a prior belief due to increased noise – as well as noise in altruistic preferences guide choices.

The Impact of Cognitive Noise Equipped with the probabilistic choice rule, I can more closely investigate the (numerical) impact of an increase in noise – both in monetary payments and altruistic preferences – on altruistic choices. Figure 1 simulates both the impact of increasing noise $\nu_{\frac{\text{self}}{\text{other}}}$ and $\nu_{\frac{\beta}{1-\beta}}$ on the probability of choosing self as a function of the ratio $\frac{\text{self}}{\text{other}}$ (equation 7) for varying values of $\mu_{\hat{r}}$. Throughout all panels, I fix $\beta = 0.3$ and $\sigma_{\hat{r}} = 1$, whereas $\nu_{\frac{\beta}{1-\beta}} = 0.25$ in the top and $\nu_{\frac{\text{self}}{\text{other}}} = 0.25$ in the bottom row.

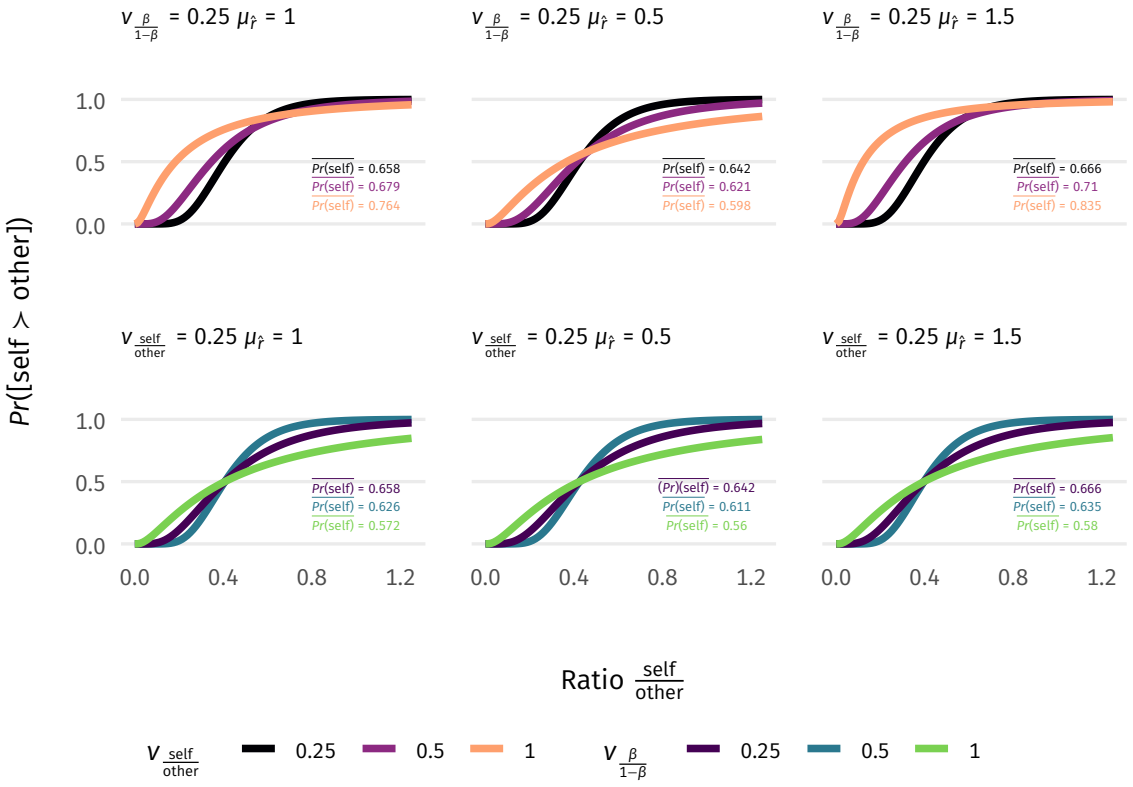


Figure 1. Impact of Cognitive Noise on Altruistic Choices This figure plots the impact of changes in cognitive noise $\nu_{\frac{\text{self}}{\text{other}}}$ (top) and $\nu_{\frac{\beta}{1-\beta}}$ (bottom) on the probabilistic choice function (equation 7) depending on different values of $\mu_{\hat{r}}$. Throughout all panels $\beta = 0.30$ and $\sigma_{\hat{r}} = 1$. The plot also includes average values of choosing self.

Consider the first row of Figure 1. There, I vary $\nu_{\frac{\text{self}}{\text{other}}} \in [0.25, 0.5, 1]$ with fixed values of $\nu_{\frac{\beta}{1-\beta}} = 0.25$ and $\mu_{\hat{r}} = 1$ (i.e., $\ln \delta = 0$). An increase in noise in perceiving monetary payments *increases* choices for self. With increased noise, the DM relies more strongly on their prior knowledge (i.e., α decreases) and $\ln \frac{\text{self}}{\text{other}}$ is attenuated towards 0. Given the log space of the choice rule, an attenuation towards 0 (of $\ln \frac{\text{self}}{\text{other}}$) implies an increase towards 1 on the original scale. In other words, smaller values of $\frac{\text{self}}{\text{other}}$ are perceived to be larger.

However, differences in the values of the prior mean will lead to different effects on choices: Consider the second graph in the first row, where $\mu_{\hat{r}} = 0.5$, i.e., an "inter-

mediate" intuitive perception of monetary payments. Now, an increase in $\nu_{\frac{\text{self}}{\text{other}}}$ *decreases* choices for self as larger values of $\frac{\text{self}}{\text{other}}$ will be downwards adjusted due to the impact of the prior. Conversely, if $\mu_{\hat{r}} = 1.5$ (third graph), an increase in $\nu_{\frac{\text{self}}{\text{other}}}$ (again) *increases* choices for self, which is quantitatively larger compared to the first instance. Overall – in this exercise – an increase in $\nu_{\frac{\text{self}}{\text{other}}}$ will therefore increase choices for self, unless $\mu_{\hat{r}} < 1$ (i.e., an "intermediate" payment perception).

In the second row, $\nu_{\frac{\text{self}}{\text{other}}}$ remains fixed at 0.25, but noise in preferences varies between $\nu_{\frac{\beta}{1-\beta}} \in [0.25, 0.5, 1]$. Focusing on the first graph, note that an increase in noise in preferences *decreases* choices for self. The origin of this effect – recall that equation 7 abstracts from the impact of a prior over preferences – lies in the log-normal noise structure of the mental signals $s_{\frac{\beta}{1-\beta}}$: Due to the concavity of the log-transform, increasing signal noise leads to stronger attenuation for larger values of the z-score of the difference between a given trials' payment ratio $\frac{\text{self}}{\text{other}}$ and the preference-induced threshold $\frac{\beta}{1-\beta}$, which in turn, translates into *fewer* choices for self. Varying $\mu_{\hat{r}}$ (second and third graph in the second row) leads to changes in the overall level of choices for self, as $\alpha < 1$, yet the effect of $\nu_{\frac{\beta}{1-\beta}}$ on choices remains largely unaffected. Note further that – across all instances – an increase in $\nu_{\frac{\beta}{1-\beta}}$ does not shift the *indifference values*, as $\nu_{\frac{\beta}{1-\beta}}$ is absent from the numerator of equation 7. The difference in average choices is thus driven by the asymmetric effect of the log-normal noise structure.[5]

Overall, the simulations show that the impact of increasing noise will depend on whether increased cognitive noise is primarily in perceiving monetary payments or altruistic preferences – circling back to the discussion in Section 2. Due to the Bayesian regularization towards a prior belief for monetary payments, the impact of noise in monetary payments will additionally depend on the moments of that prior, as illustrated.

Hypotheses Based on these discussions, I can formulate hypotheses that center on the potential mechanisms of a treatment effect due to increased noise. 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.[6] Similar to the exercise above, they should also be understood *ceteris paribus*.

---

5 Accordingly, a linear version of equation 7 with linear encoding and Gaussian priors, i.e., where $Pr(self) = \Phi\left(\frac{\alpha \times \frac{\text{self}}{\text{other}} - \frac{\beta}{1-\beta} - \delta}{\sqrt{\nu^2_{\frac{\text{self}}{\text{other}}}\alpha^2 + \nu^2_{\frac{\beta}{1-\beta}}}}\right)$ with $\delta = 1-(1-\alpha)\mu_{\hat{r}}$ does not feature such an asymmetrical effect of noise.

6 The hypotheses further illustrate the considerable degree of flexibility of the Bayesian model, which requires a careful interpretation of the empirical results and the exact mechanisms of a potential treatment effect in the experimental data later on.

Hypothesis 1_a (Noise in Payments a): An increase in $\nu_{\frac{\text{self}}{\text{other}}}$ given $\mu_{\hat{r}} \geq 1$ **increases** average choices for self.

Hypothesis 1_b (Noise in Payments b): An increase in $\nu_{\frac{\text{self}}{\text{other}}}$ given $\mu_{\hat{r}} < 1$ **decreases** average choices for self.

Hypothesis 1_c (Noise in Preferences): An increase in $\nu_{\frac{\beta}{1-\beta}}$ **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 following a "cognitive lens" to altruistic choices more generally. Throughout, I do not formulate assumptions on whether $\nu_{\frac{\text{self}}{\text{other}}}$ or $\nu_{\frac{\beta}{1-\beta}}$ is the more appropriate measure of cognitive noise in a particular instance, but understand both to be measures of different aspects of cognitive noise. I therefore refrain from discussing them separately for the additional hypotheses.

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 magnitudes, 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 be positively correlated across domains within a person. Supporting evidence in this direction is presented by Frydman and Jin (2022) and 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, noisy mental representations of problem features are 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 lists. For *altruistic choices*, this implies that an association between measures of cognitive ability and individual measures of cognitive noise $\nu_{\frac{\text{self}}{\text{other}}}$ and $\nu_{\frac{\beta}{1-\beta}}$ – key drivers of choice inconsistency and bias – should emerge, with more cognitively able people exhibiting lower values of noise.

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

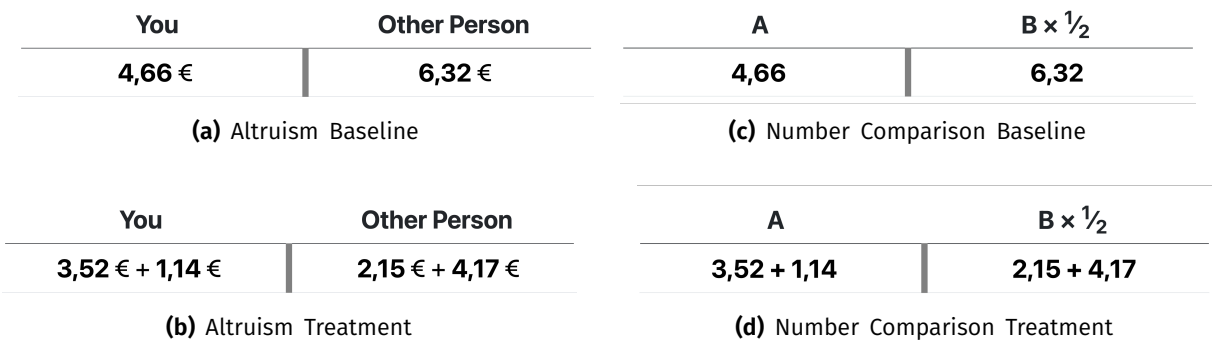
3 Experiment

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 during 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 A2.

3.1 Part 1: Altruistic Choice

Altruistic Choice In line with the theoretical setup, the experiment centers 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 choose four distinct values for other_k: 6.55 €, 9.26 €, 13.10 €, and 18.52 €[7] and calculate the indifference value $\text{self}_{indiff} \sim \frac{\beta}{1-\beta} \times \text{other}_k \ \forall \ \beta \in [0, 0.05, ..., 0.55]$ for all four values of other_k. This results in 4×12 unique combinations of self and other (see Figure A3 for an illustration). Each of these combinations is repeated five times and I call a group of five identical trials

7 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., $\sqrt{2}$.



(a) Altruism Baseline **(c)** Number Comparison Baseline

(b) Altruism Treatment **(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-calculated sums replace monetary values. **(c)** Baseline condition in the number comparison task. **(d)** Treatment condition number comparison. Participants choose using the "a" ("You"/"A") and "l" key ("Other Person"/"B") on the (German) keyboard.

a "game" throughout. 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 use payments including cent values to encourage participants to approach the decisions more approximatively.[8]

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 to 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 payoff-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 variation in Enke et al. (2023), 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 increases, 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

8 A critique of this approach could be that this leads participants to only focus on the main digit of the payments and simply ignore the cent values. While this would be 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 & Vanhuele, 2023). Note also that e.g., Dehaene and Marques (2002, p. 708) explicitly avoid round numbers in their stimuli, which are prices of different items.

signals.[9] While one could well expect this variation to impact $\nu_{\frac{\text{self}}{\text{other}}}$ and $\nu_{\frac{\beta}{1-\beta}}$ differently, I leave this as an open empirical question. In terms of the design of the variation, I choose sums as relatively simple mathematical operations to allow participants to still reasonably engage in the repeated trials of the experiment 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 systematically: I first (uniformly) draw a random number between 0 and the smaller number of the self, other pair. In the example, I drew $\text{self}_1 = 3.52$ € 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 $\text{self}_1 > \text{other}_1$, yet other > self). Furthermore, the *position* of $\text{self}_1, \text{self}_2$ and $\text{other}_1, \text{other}_2$ is randomly shuffled for each participant individually. This procedure encourages paying attention to all four components in all trials. Further, it hinders the possibility of gauging the underlying value of self or other by just focusing on the positions of the components. Table A1 provides the complete overview of all 240 trials, including the values for $\text{self}_1, \text{self}_2$ and $\text{other}_1, \text{other}_2$, 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 A12 for screenshots).

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 an objectively correct solution (A in the example) while aiming to mirror the "mechanics" (or "mental arithmetics") of the altruism decisions – comparing two numbers – 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

9 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 are 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 & Simoncelli, 2006; Weiss et al., 2002).

not to 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 & Jin, 2022; Frydman & 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.[10]

At the end of Part 2, I elicit 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 pays out at the end according to the randomized quadratic scoring rule (Hossain & Okui, 2013; Schlag & Van Der Weele, 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) pro-sociality 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 A13 shows a screenshot of CRT4. One of the six questions is drawn randomly and awards a

10 I chose to implement a time-sensitive incentivization as the task would be much more trivial to solve otherwise. 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 in 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 prevent high levels of perceived time pressure.

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 for Cognition scale (Cacioppo & Petty, 1982) developed by Beißert et al. (2014). I choose a more extensive set of cognition-related measures to compare the standard CRT questions to alternative measures related to cognitive ability.[11]

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.[12] 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 A15 for a screenshot). I also elicit survey questions related to excuse-taking (see Figure A14).[13]

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 A16 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 Preferences 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.

11 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.

12 König-Kersting (2024) does not identify differences in responses between (non-)incentivizing social norm elicitation in a large-scale experiment.

13 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.

4 Results

This section presents the empirical results, first focusing on 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 $\frac{\text{self}}{\text{other}}$ featured in the experiment, separately drawn for the Baseline and Treatment group. The Baseline data offers several insights into participants' altruistic preferences: First, unsurprisingly, the larger the payment self compared to other, the more frequently subjects choose 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 $\frac{\text{self}}{\text{other}} = 0.474$, which implies that participants roughly care twice as much about their payoff compared to the payoff of another person.

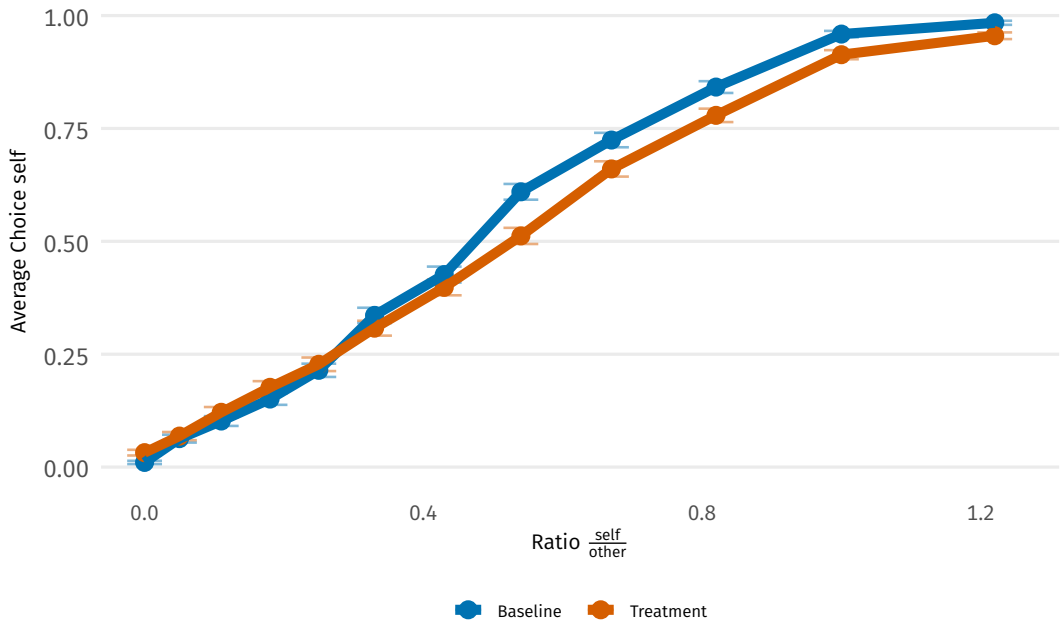


Figure 3. Altruistic Choices in Baseline and Treatment Group This plot shows the association between average choice for self and distinct values of the ratio $\frac{\text{self}}{\text{other}}$, separately drawn for the Baseline and Treatment group, with 95% confidence intervals.

This behavior largely aligns 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 the present framework as the weight a DM places on the well-being of another person (given the DM is better off).[14] Reviewing over 40 articles, Nunnari and Pozzi (2024) 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 participants here, Carpenter and Robbett (2024), Von Schenk et al. (2023) and Klockmann et al. (2022) estimate values that correspond to their subjects caring roughly twice as much about their payment compared to that of another participant.

In the Treatment condition, these statements about 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 $\frac{\text{self}}{\text{other}}$ is *flatter* compared to the Baseline condition. For small values of $\frac{\text{self}}{\text{other}}$, 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 $\frac{\text{self}}{\text{other}}$ as only 95,53 % of choices correspond to self if self $>$ other. Furthermore, 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 $\frac{\text{self}}{\text{other}}$ required for indifference in the Treatment group corresponds to 0.528, a 5.4 percentage points larger ratio compared to the Baseline. Using a linear probability model, Table A2 confirms that both the overall level of choices for self is 2.2 percentage points lower and that an increase in $\frac{\text{self}}{\text{other}}$ by 1 has a 6.7 percentage points lower effect on choices for self in the Treatment group (for a Probit model, see Table A3). Although their underlying preference should remain the same as in the Baseline, the Treatment group shows a dampened reaction to changes in incentives and chooses self significantly less frequently, i.e., behaves more altruistically.

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.*

14 Note that the framework developed here does not distinguish between being ahead and being behind as, by construction, a subject is ahead if they choose self and behind if they choose other. Thus, the present setup does not allow for separating these two motivations; instead, it comprises them into one. The fact that subjects overall substantially weigh the other person's payment could also be related to how the decision in Figure 2 is displayed, i.e., not including the 0 € consequence for either person.

Both a flatter association between varying payments and choices and a *bias* towards more altruistic choices can be rationalized in light of the theoretical framework.[15] Recall Figure 1, which outlines the impact of an increase in noise on the probability of choosing self (equation 7). The treatment effect could originate from either an increase in $\nu_{\frac{\text{self}}{\text{other}}}$ (H_{1b}) or an increase in $\nu_{\frac{\beta}{1-\beta}}$ (H_{1c}), i.e., either through an increase in "noise in payments" coupled with an additional adjustment towards an intermediate payment prior or a "mechanical" increase in "noise in preferences".

4.2 Altruistic Choices: Probabilistic Model

I now turn to probabilistic modeling to estimate the (posterior) probability of the parameter values given the experimental data and investigate the mechanisms of the treatment effect. Given the choice model formulated in equation 7, this approach allows to (i) inspect whether (an increase in) noise in payments or preferences dominates the other, (ii) infer the parameters of the prior of monetary payments, and thereby (iii) test the potential mechanisms of the treatment effect. 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, p. 4). Here, I estimate a *Bayesian Hierarchical Model* that determines the prior for the individual parameter values from the data. In hierarchical models, individual 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 by allowing (some) hyper-parameters to differ between conditions c. More specifically, the hierarchical model assumes that the individual parameter vector $\boldsymbol{\theta_i} = \left(\nu_{\frac{\text{self}}{\text{other}},i}, \nu_{\frac{\beta}{1-\beta},i}, \beta_i, \mu_{\hat{r},i}, \right)$ of individual i is – on the log-scale – drawn from a multivariate normal distribution:

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

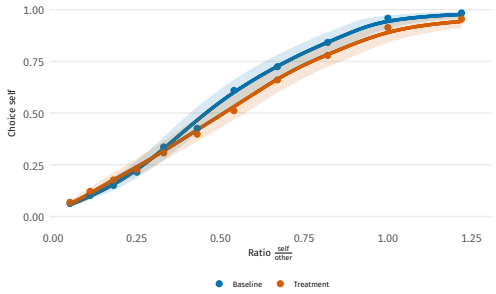
where $\boldsymbol{\mu} = (\mu_c^{\nu_{\frac{\text{self}}{\text{other}}}}, \mu_c^{\nu_{\frac{\beta}{1-\beta}}}, \mu^{\frac{\beta}{1-\beta}}, \mu^{\mu_{\hat{r}}})$ is the vector of the population-means of the parameter distributions. Note that both $\mu_c^{\nu_{\frac{\text{self}}{\text{other}}}}$ and $\mu_c^{\nu_{\frac{\beta}{1-\beta}}}$ are allowed to differ between

15 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 A9) or (ii) the treatment only working for larger numbers (Figure A9 and Table A13).

Baseline and Treatment group $\mu_B^{\nu_{\frac{self}{other}}}$, $\mu_T^{\nu_{\frac{self}{other}}}$, $\mu_B^{\nu_{\frac{\beta}{1-\beta}}}$, $\mu_T^{\nu_{\frac{\beta}{1-\beta}}}$ representing potential treatment differences along both types of cognitive noise. All other hyper-parameters remain identical across conditions. $\Sigma = \text{diag}(\tau)\Omega\text{diag}(\tau)$, where Ω is the correlation matrix of individual parameters and τ is a vector of standard deviations. Note that, without loss of generality, I set $\sigma_{\hat{r},i} = 1 \, \forall \, i$ (p. 32 Oprea & Vieider, 2024; Natenzon, 2019). The hierarchical model requires specifying prior distributions for all hyper-parameters and I choose weakly informative priors (see Section A.3.1 for details and prior predictive checks, also see Gelman et al., 2015). 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	0.316	0.316	0.008	0.3	0.332	1
Prior Mean Monetary Payments μ_r	1.049	1.005	0.304	0.533	1.652	1.01
Group Specific:						
Noise Baseline $\nu_{\frac{\beta}{1-\beta},B}$	0.186	0.185	0.022	0.143	0.23	1
Noise Treatment $\nu_{\frac{\beta}{1-\beta},T}$	0.181	0.18	0.032	0.121	0.243	1
Noise Baseline $\nu_{\frac{self}{other},B}$	0.313	0.311	0.028	0.259	0.367	1
Noise Treatment $\nu_{\frac{self}{other},T}$	0.399	0.398	0.038	0.325	0.475	1
Weight on Payments Baseline α_B	0.911	0.912	0.014	0.882	0.938	1
Weight on Payments Treatment α_T	0.862	0.863	0.023	0.817	0.905	1
Prior Threshold Baseline δ_B	1	1	0.025	0.951	1.05	1.01
Prior Threshold Treatment δ_T	1	0.999	0.039	0.925	1.076	1.01

(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 (+ 1000 warmup) per each of four chains. Parameters correspond to the 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 ($\hat{R} = 1$ indicating convergence). **(b)** Average (over individuals) and predicted choices, including 95% HDI.

The Table in panel (a) of Figure 4 summarizes the parameters of the model and (b) plots average and predicted choices (including the 95 % HDI). Given the hierarchical nature of the model, I inspect parameters on the population level, i.e., the mean of the log-normal hyper-distribution of a given parameter (instead of average individual parameters that would assign equal weight to each participant).[16] The table contains the mean, median and standard deviation of the posterior samples of the respective parameter, the 95 % credible interval, the shortest interval containing 95 % of probability mass as well as 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 $\beta = 0.316\,[0.3 - 0.332]$ aligns with the behavior described previously: on average, participants weigh approximately

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

their payment twice as important as the other person's payment. The probabilistic model further yields an estimate for the mean of the prior distribution of monetary payments $\mu_{\hat{r}} = 1.049\,[0.533 - 1.652]$, which corresponds – on average over the posterior distribution – to an "ignorance" intuition as mentioned previously, i.e., that participants intuitively do not distinguish between $\widehat{\text{self}}$ and $\widehat{\text{other}}$. However, note the large degree of uncertainty of this estimate with a 95 % probability that $\mu_{\hat{r}}$ is between 0.533 and 1.652. This will be important for discussing potential mechanisms of the treatment effect should noise in monetary payment perception be higher in the Treatment group.

This leads to the analysis of group-specific parameters. Recall that the primary goal of the chosen treatment variation was to increase noise levels in the Treatment group, but without pre-specifying if the variation would influence $\nu_{\frac{\beta}{1-\beta}}$ or $\nu_{\frac{\text{self}}{\text{other}}}$ more strongly. Regarding $\nu_{\frac{\beta}{1-\beta}}$, the probabilistic model does not indicate a difference between the Baseline and Treatment group with overall very similar levels of $\nu_{\frac{\beta}{1-\beta},B} = 0.186\,[0.143 - 0.23]$ and $\nu_{\frac{\beta}{1-\beta},T} = 0.181\,[0.121 - 0.243]$. Based on the posterior samples, I can also directly calculate probabilistic statements about a potential difference, which indicates that $P(\nu_{\frac{\beta}{1-\beta},B} < \nu_{\frac{\beta}{1-\beta},T}) = 0.434$, confirming the conclusion of no difference. This is also supported by the hyper-parameters of the preference noise distribution (on the log scale) not differing between groups with $\mu_B^{\nu_{\frac{\beta}{1-\beta}}} - \mu_T^{\nu_{\frac{\beta}{1-\beta}}} = \text{-}0.033\,[\text{-}0.371 - 0.301]$ (see Table A7). While there is, therefore no group difference, note that both $P(\nu_{\frac{\beta}{1-\beta},B} > 0) = 1$ and $P(\nu_{\frac{\beta}{1-\beta},T} > 0) = 1$. This implies that altruistic preferences are not perceived "noiselessly", yet the chosen treatment variation – encapsulating monetary payments in to-be-calculated sums – did not affect (the noise of) this perception.

This is in contrast to noise in perceiving monetary payments: Here, the probabilistic model suggests that the treatment variation did, in fact increase noise levels: $\nu_{\frac{\text{self}}{\text{other}},B} = 0.313\,[0.259 - 0.367]$, $\nu_{\frac{\text{self}}{\text{other}},T} = 0.399\,[0.325 - 0.475]$ and $\mu_T^{\nu_{\frac{\text{self}}{\text{other}}}} - \mu_B^{\nu_{\frac{\text{self}}{\text{other}}}} = 0.244\,[0.045 - 0.445]$ (see Table A7). Higher noise levels in the Treatment group translate into lower values of α, i.e., $\alpha_T = 0.862\,[0.817 - 0.905]$ and $\alpha_B = 0.911\,[0.882 - 0.938]$. Moreover, with $P(\nu_{\frac{\text{self}}{\text{other}},B} > \nu_{\frac{\beta}{1-\beta},B}) = 0.998$ and $P(\nu_{\frac{\text{self}}{\text{other}},T} > \nu_{\frac{\beta}{1-\beta},T}) = 0.999$, noise in perceiving monetary payments is also generally higher compared to noise in preferences in addition to the larger group differences.

This leads to a discussion on the origin of the treatment effect, i.e., which hypothesis is most supported by the data. Larger group differences in noise levels in perceiving monetary payments strongly suggest that the treatment variation – and, in turn, the mechanism of the observed treatment effect – operates through differences in monetary payment perception. In particular, higher levels of $\nu_{\frac{\text{self}}{\text{other}},T}$ leading to *fewer* choices for self suggest H_{1b} as a candidate hypothesis and a partial adjustment towards some "intermediate" payment perception as the driver of treatment differences. However, the probabilistic model does not conclusively support this conclusion: With the

large degree of uncertainty surrounding $\mu_{\hat{r}}$ and the fact that both $\delta_B = 1\,[0.951 - 1.05]$, $\delta_T = 1\,[0.925 - 1.076]$ and $P(\delta_T > \delta_B) = 0.496$, i.e., no group difference in the prior-induced threshold despite higher noise levels in the Treatment group, the experimental data does not provide sufficient support for a strong claim in favor of H_{1b}.

This becomes more evident following a model comparison in Figure A6. There, I compare the predictive power of the "full" model (equation 7) with various simpler variants, each one abstracting from a potential mechanism of the treatment effect.[17] While the "full" model provides the highest goodness-of-fit among all models considered, the differences compared to the simpler models are relatively minuscule.[18] Importantly, the simpler variants include a model that only incorporates noise in altruistic preferences, which performs nearly as well as the "full model". This, in turn, prohibits discarding H_{1c} entirely.[19]

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

Aside from the unclear mechanisms of the treatment effect, the probabilistic model nonetheless captures average and individual behavior well. The average predicted choice curve in panel (b) shows a strong overlap between average and predicted choices, with comparably tight HDI. In Figure A5, I further plot individual average and predicted choices with a rank-correlation of $\rho = 0.99$ between predicted and actual individual average decisions. Overall, these results thus support the general modeling approach.

4.3 Number Comparison Task: Theory and Descriptives

As stated, the evidence thus far does not conclusively inform about the origin of the treatment effect in altruistic choices. Therefore, I further investigate the mechanism of the treatment variation with the data from the number comparison task. This data is

17 More specifically, I formulate models that either set (i) the payment prior mean $\mu_{\hat{r}} = 1$ or (ii) the noise in altruistic preferences $\nu_{\frac{\beta}{1-\beta}} = 0$, or (iii) the noise in monetary payments $\nu_{\frac{\text{self}}{\text{other}}} = 0$. I also include a (iv) standard random utility benchmark for reference. 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.

18 With all variants of the "full model" outperforming the standard random-utility benchmark – supporting the overall modeling approach.

19 More concretely, this result suggests that the possibility remains that the mechanism of the treatment effect still operates through an increase in noise in preferences, yet that this increase cannot be conclusively identified and separated from an increase in noise in monetary payment perception in the "full" model. This also fits the relatively small differences in the *indifference value* between Baseline and Treatment (Figure 3).

insightful as the choice in the number comparison task share 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:

$$Pr[(A \succ B \times {}^1\!/_2] = \Phi\left(\frac{\alpha' \times \ln\left(\frac{\mathrm{A}}{\mathrm{B}}\right) - \ln(\frac{1}{2}) - \ln(\delta')}{\nu^2_{\frac{\mathrm{A}}{\mathrm{B}}}\alpha'}\right) \tag{9}$$

where $\frac{\mathrm{A}}{\mathrm{B}}$ is the ratio of numbers A and B (previously self and other), $\alpha' = \frac{1}{1+\nu^2_{\frac{\mathrm{A}}{\mathrm{B}}}}$ 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 (i.e., that the treatment impacts $\nu_{\frac{\text{self}}{\text{other}}}$ and $\nu_{\frac{\mathrm{A}}{\mathrm{B}}}$ similarly) and that this functional form is appropriate, investigating the number comparison data allows to compare between H_{1b} and H_{1c}: if H_{1b} is the driver behind the treatment effect, the treatment effect will be qualitatively similar in the number comparison task and an "intermediate" perception will lead to fewer choices for A. If, in contrast, H_{1c} is the appropriate hypothesis (and in turn, an increase in noise in preferences dominated noise in monetary payments previously), an increase in noise will *increase* choices for A, given $\mu_{\hat{r}'} \geq 1$.

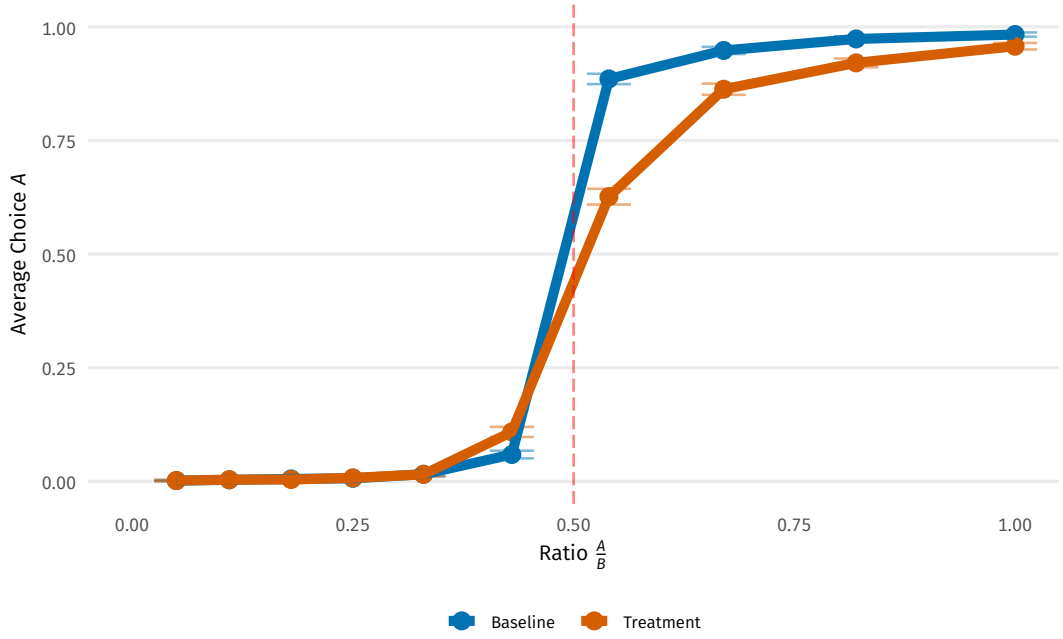


Figure 5. Number Comparison Baseline and Treatment group This plot shows the association between average choice for A and distinct values of the ratio $\frac{\mathrm{A}}{\mathrm{B}}$, separately drawn for the Baseline and Treatment group with 95% confidence intervals around mean values.

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 $\frac{\mathrm{A}}{\mathrm{B}}$, separately drawn

for Baseline and Treatment. 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 ($p < 0.001$).[20] 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 A4, 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 A5). Another apparent observation is that choices are much more *consistent* in the number comparison task than altruistic choices. This is unsurprising given that the common threshold of $^1/_2$ replaces an individual-specific preference threshold, which eliminates choice differences due to individual heterogeneity in altruistic preferences and noise in its perception.

Result 3: *In the number comparison task, The Treatment group again shows a 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 $0 < \widehat{\frac{A}{B}} < 1$ (and $0 < \frac{\widehat{self}}{\widehat{other}} < 1$), which in particular biases the perception of larger ratios downwards and turns the Treatment group towards fewer choices for A and self. I now turn to a closer inspection of the mechanism, again using a probabilistic model.

4.4 Number Comparison Task: Probabilistic Model

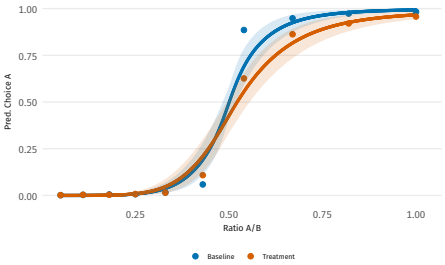
Equivalent to Section 4.2, I can estimate a probabilistic model based on the number comparison data and investigate the probability of the parameter values of equation 9. The estimated parameters are shown in panel (a) of Figure 6.

Due to the absence of altruistic preferences in the choice rule, the treatment-group invariant parameters now consist of only the mean of the prior numerical magnitude. The probabilistic model indicates that $\mu_{\hat{r}'} = 0.483\,[0.393 - 0.603]$, which corresponds to an "intermediate" intuitive perception of numerical magnitudes, that intuitively $\widehat{\frac{A}{B}} < 1$. Before interpreting the impact of this prior on choices in more detail, it

20 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 $\frac{A}{B} = 0.5$).

	mean	median	sd	hdi 2.5%	hdi 97.5%	$\hat{R}$
Base Parameters:						
Prior Mean Num. Magnitudes $\mu_{\hat{r}'}$	0.515	0.51	0.056	0.411	0.625	1
Group Specific:						
Noise Baseline $\nu_{\frac{A}{B},B}$	0.198	0.197	0.009	0.18	0.216	1
Noise Treatment $\nu_{\frac{A}{B},T}$	0.286	0.286	0.014	0.26	0.315	1
Weight on Payments Baseline α'_B	0.962	0.962	0.003	0.955	0.969	1
Weight on Payments Treatment α'_T	0.924	0.924	0.007	0.91	0.937	1
Prior Threshold Baseline δ'_B	1.026	1.026	0.004	1.017	1.034	1
Prior Threshold Treatment δ'_T	1.052	1.052	0.01	1.032	1.073	1

(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 (+ 1000 warmup) per each of four chains. Parameters correspond to the 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 ($\hat{R} = 1$ indicating convergence). **(b)** Average and predicted choices, including 95% HDI.

is useful to inspect the differences in noise(-related) parameters across groups first: Again, noise (now in perceiving numerical magnitudes) is larger in the Treatment versus the Baseline group with $\nu_{\frac{A}{B},T} = 0.286\,[0.26 - 0.315]$ and $\nu_{\frac{A}{B},B} = 0.198\,[0.18 - 0.216]$ and $P(\nu_{\frac{A}{B},T} > \nu_{\frac{A}{B},B}) = 1$. Note also that $\mu_T^{\nu_{\frac{A}{B}}} - \mu_B^{\nu_{\frac{A}{B}}} = 0.370\,[0.236 - 0.505]$ (c.f. Table A8). Due to larger noise levels in the Treatment group, the weight on payments is consequently smaller with $\alpha'_T = 0.924\,[0.91 - 0.937]$, $\alpha'_B = 0.962\,[0.955 - 0.969]$.

Comparing these parameter estimates to the values in Figure 4, several differences emerge: First, the values of $\nu_{\frac{A}{B}}$ are smaller compared to the values of $\nu_{\frac{\text{self}}{\text{other}}}$, yet the *treatment effect*, i.e., $\mu_T^{\nu_{\frac{A}{B}}} - \mu_B^{\nu_{\frac{A}{B}}}$ is larger compared to $\mu_T^{\nu_{\frac{\text{self}}{\text{other}}}} - \mu_B^{\nu_{\frac{\text{self}}{\text{other}}}}$. One possible explanation could be that noise in preferences previously counteracted some of the effects of the noise in number increase, leading to a smaller treatment effect in parameters and behavior.

The most striking difference, however, is that the mean of the numerical magnitude prior, i.e., $\mu_{\hat{r}'}$ is now much better identified (tighter HDI) and its value allows for a clear statement on the origin of the treatment effect: Based on the posterior samples, I calculate that $P(\mu_{\hat{r}'} < 1) = 1$, lending strong support to an "intermediate" intuitive numerical magnitude perception. Importantly, this translates into $P(\delta'_B > 1) = 1, P(\delta'_T > 1) = 1$, 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). This bias is also *larger* in the Treatment compared to the Baseline group given $P(\delta'_T > \delta'_B) = 1$.

Overall, the probabilistic model based on the number comparison data thus yields a much clearer indication of how the treatment effect operates, namely through a biased perception of numerical magnitudes. In particular, larger values of $\frac{A}{B}$ are being perceived as smaller under noise, which leads to a bias towards B more generally. This

more precise interpretation contrasts the more ambiguous explanations for the treatment effect discussed in Figure 4.

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

Furthermore, the individual choice curves depicted in panel (b) of Figure 6 show that the average choices are close to the HDI areas, indicating that the structural estimates reasonably recover average behavior. Figure A5 supports this with a rank-correlation of $\rho = 0.94$ between average and predicted individual choices for A. 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 that abstracts from any influence of the numerical magnitude prior, the model specification in equation 9 is superior as indicated by the $ELPD_{WAIC}$ values (see Figure A7). This strongly suggests that a single parameter $\nu_{\frac{A}{B}}$ is unable to capture the behavior of participants in the number comparison task and suggests that an additional driver, here a numerical magnitude prior with an "intermediate" mean, is at play.

Applying the findings from the number comparison data to the altruistic choices provides much stronger support for H_{1b}: Under the treatment variation, participants relied relatively more strongly on an intermediate intuitive perception that $\widehat{\text{self}} < \widehat{\text{other}}$, i.e., that the payment prior mean $0 < \mu_{\hat{r}} = \frac{\widehat{\text{self}}}{\widehat{\text{other}}} < 1$. This interpretation also fits 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 of altruistic preferences. This also matches the previous finding in Figure 4 that the treatment effect mainly increased noise in perceiving monetary payments.

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

However, a caveat to this interpretation is that it requires an explicit "logical transfer" from the number comparison to the altruism domain, as the altruism data alone did not allow to uncover the origin of the treatment effect. This is related to the *identification* of the model, which is – despite the simplifying assumption of $\sigma_{\hat{b}} \to \infty$ (i.e., no influence of a prior about altruistic preferences) not entirely given and thus may characterize a weakness of the approach. While identifiability has a different connotation in Bayesian models compared to a more classical understanding, one way to think about identifiability is the difference between the prior and posterior (parameter) distribution, i.e.,

how informative the data is (see, e.g., Xie & Carlin, 2006). Prior predictive checks in Figure A4 show that the parameter values and HDI (and corresponding behavior) based on the chosen priors differ from the posterior parameters in Figure 4. A notable exception to this is the mean of the monetary payment prior, where the posterior distribution is *wider*, again questioning the full identifiability of the model.

Related is a critique that, in turn, calls into doubt whether (modeling) noise in altruistic preferences is necessary to explain altruistic behavior in the present setting and if – instead – assuming only noise in perceiving numerical magnitudes is perhaps the more appropriate (and parsimonious) assumption. This seems especially pertinent given that the treatment effect likely operates through biased numerical magnitude perception as discussed above and that a model that excludes noise in altruistic preferences performs almost as well as the full model in the model comparison in Figure A6.

To address both critiques—that of a "mere logical transfer" and the possibility that modeling noise in altruistic preferences may be unnecessary, I now combine the datasets from both tasks. This approach allows to *jointly* estimate the parameters of equations 7 and 9. More specifically, I define a joint monetary payment and numerical magnitude noise term $\nu_{\frac{\text{self,A}}{\text{other,B}}}$ that, alongside joint $\mu_{\hat{r}}$, is estimated from both number comparison *and* altruistic choices, whereas $\nu_{\frac{\beta}{1-\beta}}$ and β are estimated from only the altruistic choice data. For the model estimation, this simply requires to include two likelihoods, again demonstrating the high flexibility of Bayesian methods.

The resulting parameter values of the combined estimation are shown in the appended Table A9. The combined model confirms the conclusions drawn previously: With $P(\mu_{\hat{r}} < 1) = 1$, $P(\delta_T > 1) = 1$ and $P(\delta_B > 1) = 1$, 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 more assertive probabilistic statements are due to much tighter posterior distributions, particularly around the prior mean, and speak in favor of increased identifiability of the jointly estimated model. The combined model further indicates – similar to above – a larger treatment effect for noise in monetary payments and numerical magnitudes than altruistic preferences (see Table A9).

One notable difference between the model based on the combined dataset and the original model is that, with $P(\nu_{\frac{\beta}{1-\beta},B} > \nu_{\frac{\text{self,A}}{\text{other,B}},B}) = 1$ and $P(\nu_{\frac{\beta}{1-\beta},T} > \nu_{\frac{\text{self,A}}{\text{other,B}},T}) = 0.997$ the combined model now indicates a *higher* level of noise in altruistic preferences. This is first evidence against the argument that altruistic preferences are perceived without noise. Further evidence against this argument is provided by the model comparison in Figure A8. There, I formulate a model that assumes $\nu_{\frac{\beta}{1-\beta}} = 0$, but now use the combined dataset for model estimation. In contrast to Figure A6, such a simpler model now performs considerably worse in explaining altruistic choices. Accounting only for

noise in numerical magnitude and monetary payment perception is thus insufficient for explaining altruistic choices. Finally, a simple linear regression in Table A6 shows that, given participant and game fixed effects (which account for individual differences in altruism and cognitive noise), the inconsistency in a given trial is significantly higher in the altruism compared to the number comparison task, underscoring the previous line of argument.

4.6 Alternative Explanations for the Treatment Effect

So far, the discussion has centered around the mechanism of the treatment effect operating through an increase in cognitive noise. 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.

I discuss several alternative explanations in Section A.3.3. 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 the results. However, treatment differences towards more altruism already materialize in the initial 10 (hypothetical) practice trials. Similarly, I find to impact of the round variable, i.e., in which trial a decision was made, on altruistic choices. 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 altruism due to how I constructed the sums of the Treatment group. Furthermore, in Section A.3.4, I estimate heterogeneous treatment effects and show that the treatment variation did not work systematically differently for participants, who, e.g., expect or hold different norms between the baseline and treatment variants of the altruism task. Most personal characteristics do not meaningfully contribute to heterogeneity in the treatment effect; if anything, the treatment effect is slightly weaker for participants scoring high on self-reported altruism and "Need for Cognition".

Overall, the best guess on the treatment effect's origin is 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 numerical 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.

4.7.1 Altruistic Choices and Number Comparison. As stated earlier, if similar cognitive processes (partly) guide altruistic choices and number comparison, 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 possible relationships. Table 1 contains the results of a linear probability model that explores if choices for self in the altruism choices correlate with future choices in the number comparison task *in the exact same trial*.[21] With multiple iterations per group of trials, I can include both participant- and game- (one game consisting of the five repetitions of a given trial) fixed effects, with the former including the treatment effect.

Table 1. Correlation of choices between tasks

	(1)	(2)	(3)
A chosen	0.048***		0.045***
	(0.008)		(0.010)
Correct Number Comparison		0.029***	0.006
		(0.009)	(0.010)
N	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
R^2	0.001	0.000	0.001

Note: 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 is consistent with the Treatment group being both more altruistic and making more errors in the number comparison task (but re-

21 As I have five repetitions of each unique trial in both tasks, I 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.

call 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).

In addition to choices, *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).

Leveraging the multivariate normal setup of both the base (Section 4.2) and combined probabilistic model (Section 4.5), I can further inspect individual correlations between the noise in monetary payments (and numerical magnitudes) and altruistic preferences independent of the treatment effects. The base model yields a high positive correlation with $\rho(\nu_{\frac{\text{self}}{\text{other}}}, \nu_{\frac{\beta}{1-\beta}}) = 0.576\,[0.295 - 0.845]$, yet the combined model a very small and even slightly negative correlation of $\rho(\nu_{\frac{\text{self,A}}{\text{other,B}}}, \nu_{\frac{\beta}{1-\beta}}) = \text{-}0.096\,[\text{-}0.218 - 0.024]$. Overall, this, therefore, only yields mixed evidence in favor of a positive association between noise across domains.

Alternatively, I can also correlate individual values (i.e., means of posterior distributions) of $\nu_{\frac{\text{self}}{\text{other}}}$ from the base model and $\nu_{\frac{\text{A}}{\text{B}}}$ from the number comparison model (Section 4.4). The overall correlation between these values is $\rho = 0.155$, $p = 0.007$, which indicates a small positive association between the two. If I separate the data by treatment group, I obtain $\rho = 0.245$, $p = 0.002$ for the Treatment and $\rho = 0.059$, $p = 0.470$ for the Baseline group, which suggests that noise across tasks is positively related only within the Treatment variant of the task. Notably, the Treatment correlation coefficient of 0.245 is very 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 here 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] 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.

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

Overall, I evaluate the presented evidence as tentative support for H_2 and 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 between altruistic choices and number comparison.*

4.7.2 (Self-reported) Cognitive Ability and Individual Measures of Noise. I now investigate H_3, i.e., 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 and self-reported math abilities with individual structural measures of noise, i.e., both $\nu_{\frac{\text{self}}{\text{other}},i}$ and $\nu_{\frac{\beta}{1-\beta},i}$ from the base model (Section 4.2). 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:

Table 2. Correlation structural parameters and individual characteristics: Cognition and altruism

	Noise Altr. Preference $\nu_{\frac{\beta}{1-\beta},i}$	Noise Monetary Payments $\nu_{\frac{\text{self}}{\text{other}},i}$	Altr. Preference β_i
Cognition-related:			
No. Correct CRT	-0.327***	-0.274***	0.126*
Berlin Numeracy Test	-0.247***	-0.276***	0.173**
'I am good at math'	-0.197***	-0.149**	0.059
Avg. Need for Cognition	-0.150**	-0.163**	0.125*
Avg. Deliberation	0.098	0.150**	-0.142*
Avg. Intuition	0.066	0.119*	-0.097
Altruism-related:			
Dictator Game Other	-0.120*	-0.357***	0.495***
GPS Donation	0.063	-0.049	0.146*
GPS Value Gift	-0.048	-0.107	0.137*

Note: Need for Cognition, Deliberation and Intuition averaged values. Self-reported math abilities are elicited on a 0-10 scale. Individual parameter estimates taken from model in Section 4.2. 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 Test and individual measures of both $\nu_{\frac{\beta}{1-\beta},i}$ and $\nu_{\frac{\text{self}}{\text{other}},i}$. Similarly, the higher the score on the Berlin Numeracy Test, the higher self-reported math capabilities and "Need for Cognition", the lower the individual estimate of both noise terms. These associations – though only correlational – underscore an

important point: The proposed theoretical model and the $\nu_{\frac{\beta}{1-\beta},i}$ and $\nu_{\frac{\text{self}}{\text{other}},i}$ parameters, 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 in line with the above-mentioned work that shows how CRT performance correlates with biases and mistakes in choices (Assenza et al., 2019; Augenblick et al., 2022; Chew et al., 2022; Oprea, 2024). In contrast, self-reported preferences for deliberation and intuition do not meaningfully correlate with $\nu_{\frac{\beta}{1-\beta},i}$ and only to a slight extent with $\nu_{\frac{\text{self}}{\text{other}},i}$.

Result 6: *Individual measures of cognitive noise negatively correlate with cognitive ability as measured by performance in the Cognitive Reflection Test 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 much lesser extent, with the hypothetical donation and gift-giving decision from the GPS. Furthermore, β_i positively correlates with the CRT and BNT performance[23], whereas $\nu_{\frac{\beta}{1-\beta},i}$ and even more so $\nu_{\frac{\text{self}}{\text{other}},i}$ 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: Both the correlation between noise in monetary payments and altruistic preferences, $\rho(\nu_{\frac{\text{self}}{\text{other}}}, \frac{\beta}{1-\beta}) =$ -0.277 [-0.519 − -0.034], as well as noise in altruistic preferences and altruistic preferences $\rho(\nu_{\frac{\beta}{1-\beta}}, \frac{\beta}{1-\beta}) =$ -0.768 [-0.868 − -0.661] are negatively correlated, which in turn could explain the above-mentioned patterns.

4.8 Response Times and Metacognition

I now turn to study two core components insightful for choice processes: response times and measures of "metacognition," which – as understood here – comprise several measures of 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. RT is a highly informative variable of the choice process in psychology and cognitive science (see e.g., Luce, 1991) with 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

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

for physical stimuli, such as the brightness of two lights (see, e.g., Pins & Bonnet, 1996, "Pierons Law"), as well as for numerical stimuli, such as two Arabic numerals (see, e.g., Moyer & 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 & Garagnani, 2022). RT has also been used to investigate social preferences, especially under the umbrella of dual-process models with fast (slow) decisions usually attributed to intuitive (deliberate) reasoning. Time pressure studies concluded that people intuitively tend towards cooperation (Rand & Kraft-Todd, 2014; Rand et al., 2012) ("Social Heuristics Hypothesis"), and that "fairness is intuitive" (Cappelen et al., 2016).[24]

In the present setting, I analyze RT and its correlation with behavior from various angles. First, a straightforward test is to investigate if the treatment variation, aimed at increasing the cognitive difficulty, actually leads to higher RT in the Treatment group. This is the case: On average, participants in the Baseline group took 1.33 seconds to decide on 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 to-be-calculated 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 in 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 ($\widetilde{\frac{\text{self}}{\text{other}}}_i$), resp. A and B ($\widetilde{\frac{\text{A}}{\text{B}}}_i$). I can then calculate the difference of the ratio of a current trial j to that indifference value ($\Delta\widetilde{\frac{\text{self}}{\text{other}}}_{ij} = \frac{\text{self}}{\text{other}}_j - \widetilde{\frac{\text{self}}{\text{other}}}_i$) and investigate how the RT of participant i in trial j relates to that difference.

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 $\Delta\widetilde{\frac{\text{self}}{\text{other}}}$ and $\Delta\widetilde{\frac{\text{A}}{\text{B}}}$ and scale the size of the data points proportional to their relative frequency. The polynomials are fitted to the non-aggregated

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).

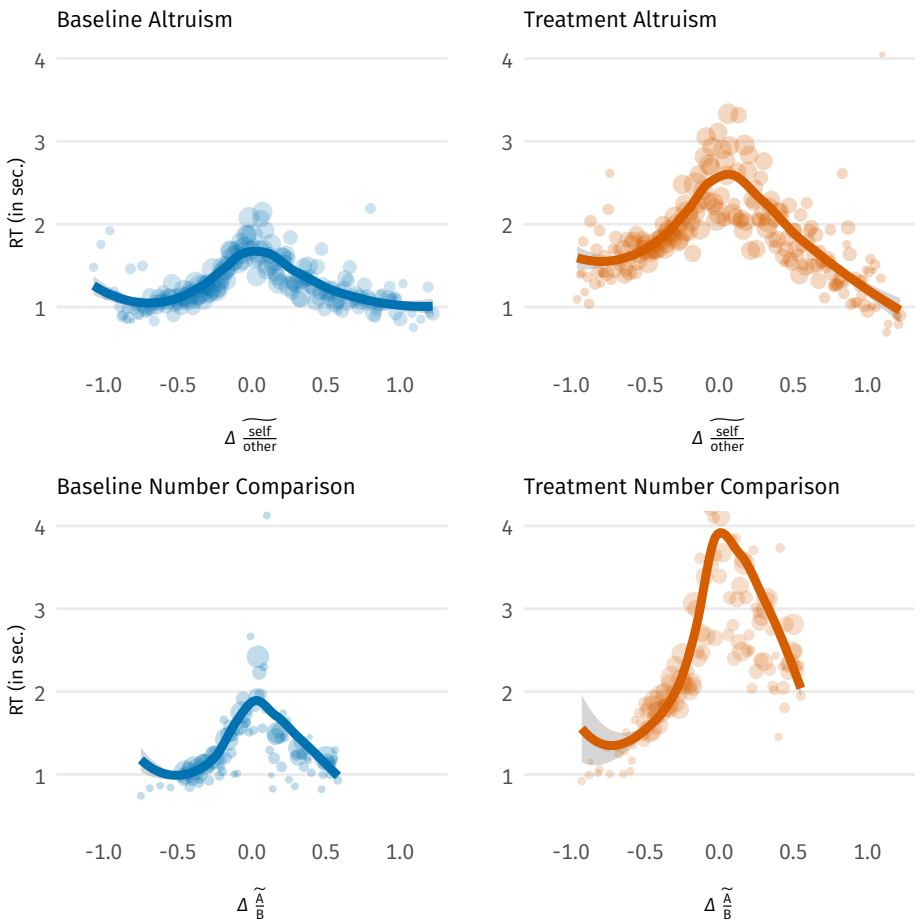


Figure 7. Distribution of RT and Distance to Predicted Indifference Ratio $\Delta\widetilde{\frac{self}{other}}_{ij} = \frac{self}{other}_j - \widetilde{\frac{self}{other}}_i$, where $\widetilde{\frac{self}{other}}_i = \widetilde{\frac{self}{other}}_i : Pr(self_i, \widetilde{\frac{self}{other}}_i) = 0.5$ and $\widetilde{\frac{A}{B}}$ and $\Delta\widetilde{\frac{A}{B}}$ 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.

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 consider this as validating evidence that the probabilistic model and the proposed decision rule in Section 2 (and Section 4.3) 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 the present 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", i.e., quick, 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.

Table 3. Correlation of RT with behavior

	(1)	(2)	(3)	(4)
$\frac{\text{self}}{\text{other}}, \frac{\text{A}}{\text{B}}$	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
N	72000	72000	60000	60000
Clustered Standard Errors	Yes	Yes	Yes	Yes
Unique Obs	300	300	300	300

Note: 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: *In the number comparison task, 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" Enke and Graeber (2023, p. 2021) 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 self-reported measures of confidence (similar to the inverse of 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.

Table 4. Treatment effects metacognition

	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:			
Avg. Attention	0.761	0.705	0.007***
Precision	0.568	0.447	<0.001***
|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*

Note: 7 participants are omitted, where $|\Delta \text{ Belief Time Spent}| > 10$ 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 the Baseline and Treatment group, alongside the p-value of a two-sided t-test. The first set of measures in the table contains the negatively-recoded self-reported confidence (the inverse of 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 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. Overall, there is no evidence of a treatment effect in the metacognitive measures.

This is markedly different in the number comparison domain: here, participants in the Treatment group report lower average attention (Baseline: 0.761, Treatment 0.705, $p < 0.01$) and precision (Baseline: 0.568, Treatment 0.447, $p < 0.01$). In addition to these self-reports, the number comparison task offers several additional measures, which all show clear group differences: in the Treatment group, participants deviate more strongly in their beliefs from their true performance[27], i.e., have a larger $|\Delta$ 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., $p = 0.01$) and again report lower confidence in these estimates (Baseline: 0.638, Treatment 0.593, $p = 0.093$). Similarly, correlations between choices, RT, metacognitive measures and choice inconsistencies are overall more pronounced in the number comparison domain compared to the altruism choices (see Table A14 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, the correlation between behavior and metacognition is more pronounced for number comparison than for altruism.*

5 Discussion of Results and Next Steps

In this paper, I established the following main results: (i) Encasing monetary payments in to-be-calculated sums causes more altruistic choices in the a simple give vs. take task. This effect most likely operates through the perception of monetary payment values, as the effect manifests comparably in the number comparison task. (ii) I observe

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

27 See also Figure A11, 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.

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 however both reacts more strongly to the treatment variation and is 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 demonstrates how an intermediate intuition of $\widehat{\text{A}} < \widehat{\text{B}}$ (and $\widehat{\text{self}} < \widehat{\text{other}}$) 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 the altruism task carry such simple (and easy to grasp) "rules", it consequently also fits to the statistical environment of the tasks of the experiment as the empirical average ratio in the average trial amounts to $\frac{\bar{\text{self}}}{\text{other}} = 0.466$ and $\frac{\bar{\text{A}}}{\bar{\text{B}}} = 0.437$. While it remains possible that an *adaptation* to the statistics of both tasks over the course of the experiments explains the origin of the treatment effect, the early emergence of treatment differences (in the practice trials) and the absence of strong evidence for learning effects (see Section 4.6) make a quick, intuitive grasp of the task a more likely explanation.

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 the present tasks, other tasks 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 relates to a point in Oprea and Vieider (2024, p. 33) who explicitly discuss differences between "naive" versus more sophisticated decision-makers when specifying a prior mean parameter. Here, I provide strong support for the presence of such non-naive intuitions. This has important implications: If people quickly grasp the "rules" or statistics of typical tasks, this potentially alters the direction of an increase 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 that identifies how behavior actually reacts to an increase in noise, which

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 more convincing.

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.

in turn depends on the prior (mean). The to-be-calculated sums proposed here are a candidate for such a causal test.

However, I fully acknowledge that the interpretation of an intuition reminiscent of the "rules" of the task is so far purely speculative and not the result of an empirical test. Future work could therefore investigate the drivers of 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) and Polanía et al. (2019) or Prat-Carrabin & Gershman, 2024) or explicitly model the noisy learning process (see Poggi (2021) for a start).

In addition, while I *theoretically* demonstrate how noise in perceiving altruistic preferences can affect choices, the implemented to-be-calculated sums likely operate through monetary payment and numerical magnitude perception. An important next step would thus be to either design and implement a variation that exclusively affects the noise in the perception of altruistic preferences without affecting numerical perception. Another possible next step is to compare the effects of the present treatment to more standard time pressure or cognitive load treatments, which potentially could also aid in better-identifying parameters of a more extensive theoretical model (that, e.g., includes prior beliefs over preferences).

(ii) Correlation of Altruism and Number Comparison The second set of results shows a (moderate) correlation between behavior and measures of noise in altruistic choices and number comparison: I both observe an increase in choices for self if a person chooses A in a future "twin"-trial and also if this person identifies 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 the present setting.[30] Nonetheless, common to both domains is the necessity to *compare*, which in turn could be related to common cognitive processes. Note that 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

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 & Garagnani, 2022, p. 313).

of a chosen setting determine this relationship, remains a largely open question and seems worthy of future investigation.[31]

(iii) Cognition and Altruistic Preferences The third result shows a correlation between measures of cognitive noise and cognitive ability. Cognitive processes are thus likely to play a role at *expressing* one's (subjective) preferences, yet a directional association with different levels of altruism is less clear (recall that I did not observe any correlation between measures of cognitive noise and altruism per se). Similar to the present 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. (2023) 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, it speaks in favor of a dampened expression of selfish preferences in the Treatment group. A similar point is raised by Enke (2024) – in light of discussions on the link between confusion and public goods contribution – in that information-processing constraints impact the translation of social preferences into behavior. Similarly, Bao and Pei (2024) interpret cognitive uncertainty as a complementary driver to social preferences in public goods contributions (see also the public goods game results in Enke et al. (2024)). What this paper adds to this discussion is that sheds closer light on the *mechanism* of the dampened expression, namely through numerical magnitude perception in the present setting, and uncovering precise estimates on the location of the prior – in the absence of a clear default option – that dictates how higher noise leads to differences in behavior. At the same time, self-reported "meta-cognitive" measures seem less relevant for explaining altruistic choice in the present setting (see iv below).

Ultimately, a dampened expression as the mechanism also provides 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 & Zinman, 2023), especially for associations between social preferences and cognitive ability (Chen et al., 2013; Hauge et al., 2009; Ponti & Rodriguez-Lara, 2015), which mostly focus on associations between the level of preferences and cognitive ability thus far. In line with the interpretation put forward here, Olschewski et

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.

al. (2018) find that in ultimatum game choices, cognitive load mostly increases choice variability only and does not impact preferences per se.[32]

(iv) RT and Metacognition The fourth set of results is related to the link between behavior, RT and metacognition. The main result 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 is *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, see 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 remain 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.

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, i.e., choosing between taking an amount self or giving an amount other. Crucially, the experiment featured a between-subject manipulation of the cognitive difficulty of choosing in the Treatment group, which was shown to-be-calculated sums instead of plain monetary values. I observe both a flatter association between changes in payments and choices as well as overall more altruistic choices in the Treatment group. After the altruistic choices, I repeated the trials of the experiment in a comparable number comparison task, where

32 See also an ongoing discussion regarding risk and time preferences and whether cognitive ability is related to choice mistakes "only" or preferences (Amador-Hidalgo et al., 2021; Olschewski et al., 2023).

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, cognitive ability and altruistic choice.

The expression of altruistic preferences – and social preferences more generally – is thus not immune to the cognitive difficulty of their implementation. This further implies that at least part of observed pro-social choices are due to (individual differences in) cognitive noise, which in turn may be related to cognitive ability. This also 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 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 related both to the fact that altruism choices as operationalized here remained relatively simple and that 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 a promising avenue for future research.

A Appendix

A.1 Theory

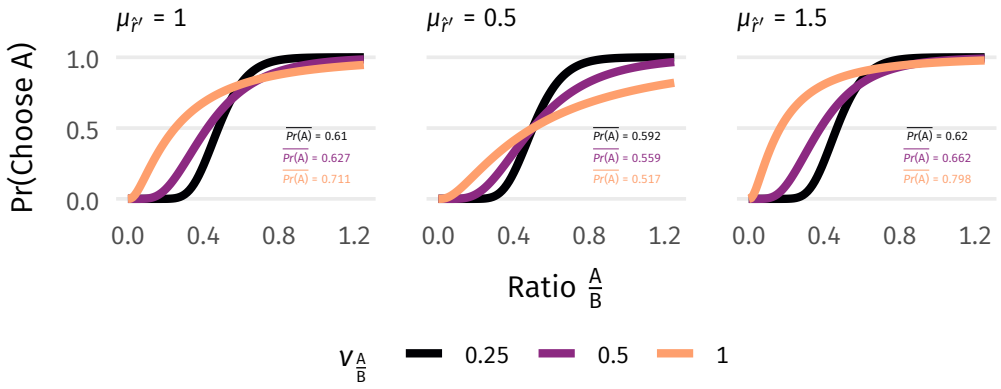


Figure A1. Impact of Cognitive Noise on Number Comparison This shows the impact of changes in noise $v_{\frac{A}{B}}$ on the choice function 9 depending on different values of $\mu_{\hat{r}'}$. Note that $\sigma_{\hat{r}'} = 1$.

A.2 Graphs and Figures Experiment

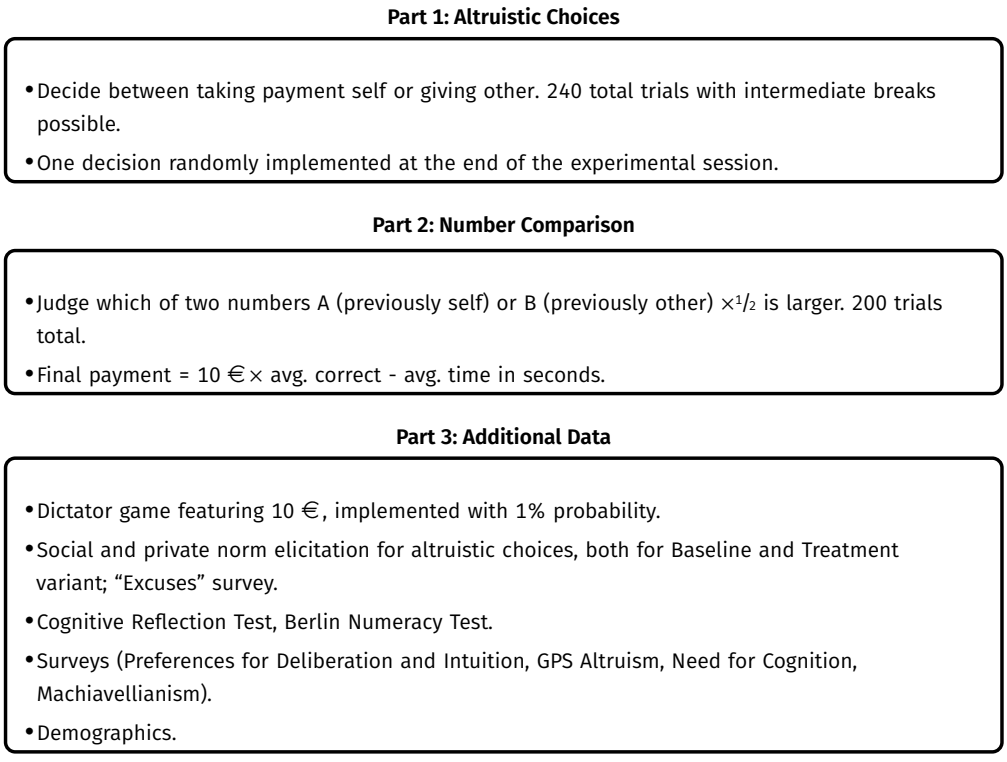


Figure A2. Graphical Outline of an Experimental Session

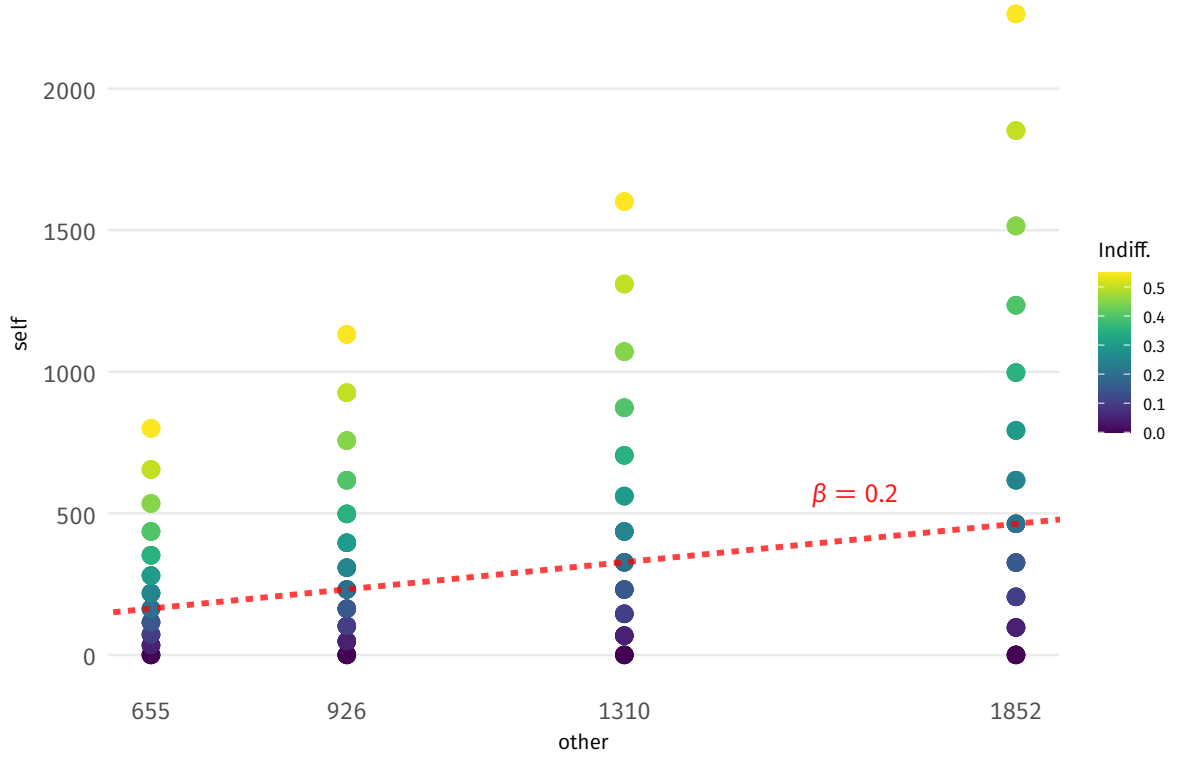


Figure A3. Payment Combinations in Altruistic Choice Task This graph shows the 48 unique combinations of self and other in the experimental trials. Each combination is repeated five times, totaling 240 decisions, with one decision randomly implemented. Note that I instructed participants precisely like 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 always decides for other in the trials below and for self above this line. Payments are in Eurocents.

Table A1. Overview of 240 trials of the 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
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

Continued on next page

Table A1 – continued from previous page

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
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
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

Continued on next page

Table A1 – continued from previous page

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
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
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

Continued on next page

Table A1 – continued from previous page

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
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
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

Continued on next page

Table A1 – continued from previous page

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
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
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

Continued on next page

Table A1 – continued from previous page

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
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
42	4	1852	793	245	548	142	1710
43	4	1852	997	239	758	1188	664

Continued on next page

Table A1 – continued from previous page

Game Identifiers		Payments		Components of Sums (Treatment)			
Game ID	Game Group	Other	Self	Self 1	Self 2	Other 1	Other 2
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

Table A2. Altruistic choice treatment effect regression

	(1)	(2)	(3)	(4)
Treatment Group	-0.022***	-0.022	0.009	-0.014
	(0.004)	(0.024)	(0.030)	(0.023)
Ratio $\frac{\text{self}}{\text{other}}$			0.900***	0.867***
			(0.018)	(0.013)
Treatment Group * Ratio $\frac{\text{self}}{\text{other}}$			-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
N	72000	72000	72000	72000
Unique Obs	300	300	300	300
R^2	0.001	0.001	0.431	0.517

Note: Linear probability model with choice for self as dependent variable. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky & Tipton, 2018)) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table A3. Altruistic choice treatment effect regression probit model

	(1)	(2)	(3)	(4)
Treatment Group	-0.057***	-0.057	0.154	0.663***
	(0.009)	(0.060)	(0.137)	(0.171)
Ratio $\frac{\text{self}}{\text{other}}$			3.380***	6.225***
			(0.153)	(0.065)
Treatment Group × Ratio $\frac{\text{self}}{\text{other}}$			-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 R^2	0	0	0.375	-
N	72000	72000	72000	72000

Note: Probit Model with choice for self as dependent variable. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky & Tipton, 2018)) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table A4. Number comparison treatment effect regression

	(1)	(2)	(3)
Treatment Group	-0.037***	-0.220***	-0.220***
	(0.005)	(0.022)	(0.022)
Ratio $\frac{A}{B}$		-0.932***	-0.932***
		(0.005)	(0.005)
Treatment Group × Ratio $\frac{A}{B}$		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
N	60000	60000	60000
Unique Obs	300	300	300
R^2	0.001	0.794	0.798

Note: Linear probability model with choice for A as dependent variable. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky & Tipton, 2018)) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table A5. Number comparison treatment effect probit model

	(1)	(2)	(3)	(4)
Treatment Group	-0.099***	-0.099***	0.839***	0.800***
	(0.010)	(0.012)	(0.286)	(0.077)
Ratio $\frac{A}{B}$			8.693***	9.131***
			(0.576)	(0.104)
Treatment Group × Ratio $\frac{A}{B}$			-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
N	300	300	300	300
pseudo R^2	0.001	0.001	0.696	-
Num.Obs.	60000	60000	60000	60000

Note: Probit model with choice for A as dependent variable. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky & Tipton, 2018)) in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table A6. Inconsistencies across tasks regression

	(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)
N	26400	26400	26400
R^2	0.004	0.070	0.191
Participant FE	No	Yes	Yes
Game FE	No	No	Yes
Clustered Standard Errors	No	Yes	Yes

Note: 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

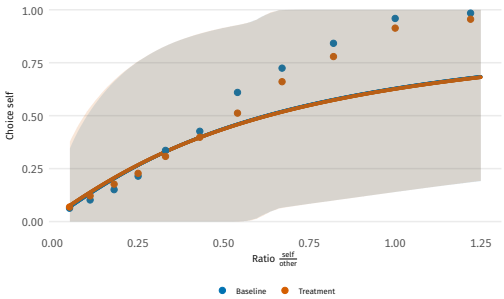
$$
\begin{aligned}
\mu^{\nu_{\frac{\text{self}}{\text{other}}}} &\sim \mathcal{N}(-0.5, 0.25) & \sigma^{\nu_{\frac{\text{self}}{\text{other}}}} &\sim \mathcal{N}^{+}(0, 0.25) \\
\mu^{\nu_{\frac{\beta}{1-\beta}}} &\sim \mathcal{N}(-0.5, 0.25) & \sigma^{\nu_{\frac{\beta}{1-\beta}}} &\sim \mathcal{N}^{+}(0, 0.25) \\
\mu^{\frac{\beta}{1-\beta}} &\sim \mathcal{N}(-0.5, 0.25) & \sigma^{\frac{\beta}{1-\beta}} &\sim \mathcal{N}^{+}(0, 0.25) \\
\mu^{\mu_{\hat{r}}} &\sim \mathcal{N}(-0.5, 0.25) & \sigma^{\mu_{\hat{r}}} &\sim \mathcal{N}^{+}(0, 0.25) \\
\mu_B^{\nu_{\frac{\text{self}}{\text{other}}}} - \mu_T^{\nu_{\frac{\text{self}}{\text{other}}}} &\sim \mathcal{N}(0, 0.25) & \mu_B^{\nu_{\frac{\beta}{1-\beta}}} - \mu_T^{\nu_{\frac{\beta}{1-\beta}}} &\sim \mathcal{N}(0, 0.25)
\end{aligned}
$$

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

Prior Summary and Predictive Checks

	mean	median	sd	hdi 2.5%	hdi 97.5%
Base Parameters:					
Altr. Preference β	0.395	0.392	0.111	0.197	0.619
Prior Mean Outcomes $\mu^{\hat{r}}$	0.65	0.619	0.175	0.34	1.004
Group Specific:					
Noise Baseline $\nu_{\frac{\beta}{1-\beta},B}$	0.634	0.616	0.159	0.34	0.911
Noise Treatment $\nu_{\frac{\beta}{1-\beta},T}$	0.653	0.617	0.229	0.279	1.104
Noise Baseline $\nu_{\frac{\text{self}}{\text{other}},B}$	0.638	0.613	0.169	0.361	1.012
Noise Treatment $\nu_{\frac{\text{self}}{\text{other}},T}$	0.653	0.612	0.24	0.289	1.116
Weight on Payments Baseline α_B	0.713	0.727	0.101	0.482	0.872
Weight on Payments Treatment α_T	0.708	0.728	0.135	0.445	0.923
Prior Threshold Baseline δ_B	1.148	1.133	0.111	0.969	1.38
Prior Threshold Treatment δ_T	1.153	1.127	0.131	0.969	1.443

(a) Prior parameter summary



(b) Average and prior-predicted choices

Figure A4. Summary Prior Probabilistic Model Altruistic Choices **(a)** Prior parameter values of equation 7 based on 10000 prior samples (i.e., before providing experimental data). Parameters correspond to the mean of log-normal hyper-distributions. Mean, median and sd mean, median and standard deviation 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.

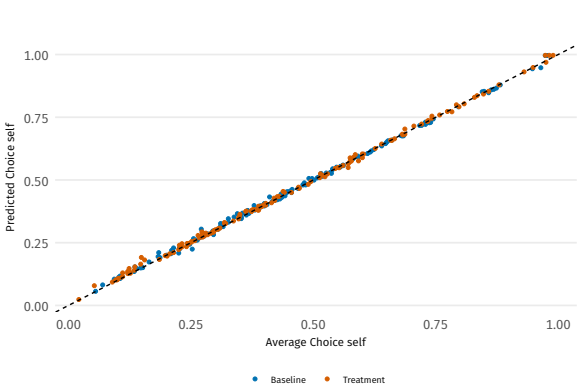
Posterior Summaries

Table A7. Posterior parameter summary average individual parameters altruistic choice

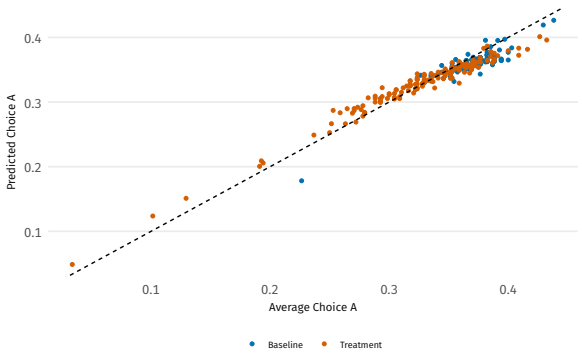
	mean	median	hdi 2.5%	hdi 97.5%	$\hat{R}$
$\mu_B^{v_{\frac{\beta}{1-\beta}}} - \mu_T^{v_{\frac{\beta}{1-\beta}}}$	-0.033	-0.028	-0.371	0.301	1.00
$\mu_T^{v_{\frac{self}{other}}} - \mu_B^{v_{\frac{self}{other}}}$	0.244	0.245	0.045	0.445	1.00
$\mu^{v_{\frac{\beta}{1-\beta}}}$	-1.783	-1.778	-2.019	-1.554	1.01
$\mu^{v_{\frac{self}{other}}}$	-1.342	-1.339	-1.523	-1.165	1.00
$\mu^{\mu^{\hat{r}}}$	-0.316	-0.316	-0.779	0.131	1.01
$\mu^{\frac{\beta}{1-\beta}}$	-0.795	-0.795	-0.875	-0.715	1.00
$\sigma^{v_{\frac{\beta}{1-\beta}}}$	0.415	0.405	0.230	0.621	1.00
$\sigma^{v_{\frac{self}{other}}}$	0.586	0.580	0.434	0.748	1.00
$\sigma^{\frac{\beta}{1-\beta}}$	0.206	0.204	0.151	0.265	1.00
$\sigma^{\mu_{\hat{r}}}$	0.786	0.781	0.447	1.145	1.00

Table A8. Posterior parameter summary hyper-parameters number comparison

	mean	median	hdi 2.5%	hdi 97.5%	$\hat{R}$
$\mu^{v_{\frac{A}{B}}}$	-1.754	-1.754	-1.844	-1.664	1
$\mu^{\mu_{\hat{r}}}$	-1.101	-1.102	-1.269	-0.938	1
$\mu_T^{v_{\frac{A}{B}}} - \mu_B^{v_{\frac{A}{B}}}$	0.370	0.371	0.236	0.505	1
$\sigma^{v_{\frac{A}{B}}}$	0.513	0.512	0.465	0.562	1
$\sigma^{\mu_{\hat{r}}}$	0.927	0.925	0.800	1.062	1



(a) Altruism Average and Individually Predicted



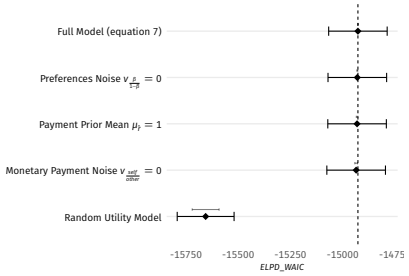
(b) Number Comparison Average and Individually Predicted

Figure A5. Individual Average and Predicted Behavior: Altruism and Number Comparison Correlation between average choice for self (a) and A (b) and predicted choice at $\frac{self}{other}$ and $\frac{A}{B}$ implemented in the experiment. Rank-correlations are $\rho = 0.999$ (a) and 0.939 (b).

Model Comparisons

Model	Choice Function	$ELPD_{WAIC}$
Full Model (equation 7)	$Pr([\text{self} > \text{other}]) = \Phi(\frac{\alpha \times \ln(\frac{\text{self}}{\text{other}}) - \ln(\frac{\beta}{1-\beta}) - \ln(\delta)}{\sqrt{v^2_{\frac{\text{self}}{\text{other}}} \alpha^2 + v^2_{\frac{\beta}{1-\beta}}}})$	-14,927.47
Payment Prior Mean $\mu_{\hat{r}} = 1$	$Pr([\text{self} > \text{other}]) = \Phi(\frac{\alpha \times \ln(\frac{\text{self}}{\text{other}}) - \ln(\frac{\beta}{1-\beta})}{\sqrt{v^2_{\frac{\text{self}}{\text{other}}} \alpha^2 + v^2_{\frac{\beta}{1-\beta}}}})$	-14,931.52
Preference Noise $v_{\frac{\beta}{1-\beta}} = 0$	$Pr([\text{self} > \text{other}]) = \Phi(\frac{\alpha \times \ln(\frac{\text{self}}{\text{other}}) - \ln(\frac{\beta}{1-\beta})}{\alpha \times v_{\frac{\text{self}}{\text{other}}}})$	-14,930.33
Monetary Payment Noise $v_{\frac{\text{self}}{\text{other}}} = 0$	$Pr([\text{self} > \text{other}]) = \Phi(\frac{\ln(\frac{\text{self}}{\text{other}}) - \ln(\frac{\beta}{1-\beta})}{v_{\frac{\beta}{1-\beta}}})$	-14,935.94
Random Utility	$Pr([\text{self} > \text{other}]) = \frac{e^{\alpha(1-\beta)\text{self}}}{e^{\alpha(1-\beta)\text{self}} + e^{\alpha\beta\text{other}}}$	-15,656.30

(a) Models Altruistic Choice

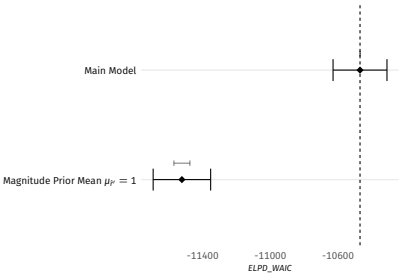


(b) $ELPD_{WAIC}$ values

Figure A6. 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 a 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	$ELPD_{WAIC}$
Main Model (equation 9)	$Pr([\text{A} > \text{B} \times {}^1/_2]) = \Phi \frac{\alpha' \log \frac{\text{A}}{\text{B}} - \log \frac{1}{2} - \log \delta'}{v' \times \alpha'}$	10,470.85
Magnitude Prior Mean $\mu_{\hat{r}'} = 1$	$(Pr[(\text{A} > \text{B} \times {}^1/_2]) = \Phi \frac{\alpha' \log \frac{\text{A}}{\text{B}} - \log \frac{1}{2}}{v' \times \alpha'}$	-11,523.44

(a) Models Number Comparison



(b) $ELPD_{WAIC}$ values

Figure A7. Model Comparison Number Comparison $ELPD_{WAIC}$ refers to the expected log predictive density as based on the widely-applicable information criterion (WAIC); A larger $ELPD_{WAIC}$ indicates a 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).

Table A9. Parameter summary combined estimation

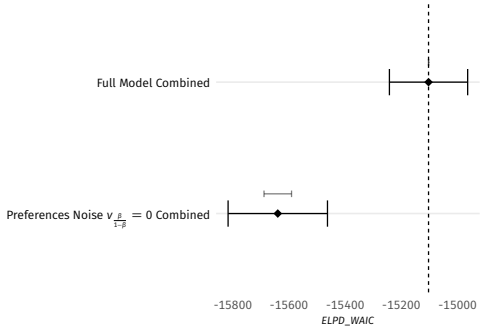
	mean	median	sd	hdi 2.5%	hdi 97.5%	$\hat{R}$
Base Parameters:						
Altr. Preference β	0.315	0.315	0.012	0.292	0.339	1
Prior Mean Outcomes $\mu^{\hat{r}}$	0.474	0.467	0.062	0.366	0.6	1
Group Specific:						
Noise Baseline $v_{\frac{\beta}{1-\beta},B}$	0.339	0.335	0.038	0.27	0.415	1
Noise Treatment $v_{\frac{\beta}{1-\beta},T}$	0.323	0.32	0.035	0.262	0.394	1
Noise Baseline $v_{\frac{\text{self,A}}{\text{other,B}},B}$	0.172	0.172	0.008	0.158	0.188	1
Noise Treatment $v_{\frac{\text{self,A}}{\text{other,B}},T}$	0.255	0.255	0.012	0.233	0.279	1
Weight on Payments Baseline α_B	0.971	0.971	0.003	0.966	0.976	1
Weight on Payments Treatment α_T	0.939	0.939	0.005	0.928	0.949	1
Prior Threshold Baseline δ_B	1.022	1.022	0.004	1.014	1.03	1
Prior Threshold Treatment δ_T	1.047	1.048	0.009	1.029	1.066	1

Table A10. Posterior parameter summary hyper-parameters combined estimation

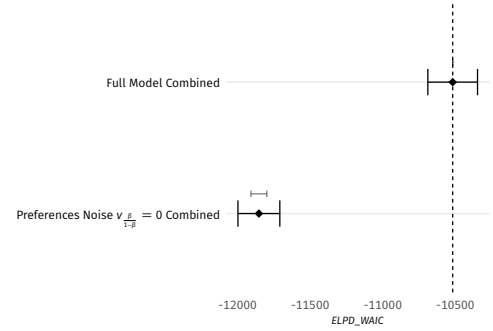
	mean	median	hdi 2.5%	hdi 97.5%	$\hat{R}$
$\mu_B^{v_{\frac{\beta}{1-\beta}}} - \mu_T^{v_{\frac{\beta}{1-\beta}}}$	-0.047	-0.046	-0.251	0.153	1.00
$\mu_T^{v_{\frac{\text{self,A}}{\text{other,B}}}} - \mu_B^{v_{\frac{\text{self,A}}{\text{other,B}}}}$	0.394	0.396	0.260	0.520	1.01
$\mu^{v_{\frac{\beta}{1-\beta}}}$	-1.519	-1.519	-1.667	-1.366	1.00
$\mu^{v_{\frac{\text{self,A}}{\text{other,B}}}}$	-1.792	-1.793	-1.878	-1.703	1.00
$\mu^{\mu_{\hat{r}}}$	-1.132	-1.133	-1.297	-0.967	1.00
$\mu^{\frac{\beta}{1-\beta}}$	-0.980	-0.980	-1.072	-0.894	1.00
$\sigma^{v_{\frac{\beta}{1-\beta}}}$	0.924	0.920	0.752	1.109	1.00
$\sigma^{v_{\frac{\text{self,A}}{\text{other,B}}}}$	0.254	0.252	0.206	0.302	1.00
$\sigma^{\frac{\beta}{1-\beta}}$	0.638	0.635	0.543	0.745	1.00
$\sigma^{\mu_{\hat{r}}}$	0.863	0.859	0.659	1.079	1.00

Model	Combined Choice Functions	$ELPD_{WAIC,A}$	$ELPD_{WAIC,NC}$
Full Model Combined	$Pr([\text{self} \succ \text{other}]) = \Phi\left(\frac{\alpha \times \ln\left(\frac{self}{other}\right) - \ln\left(\frac{\beta}{1-\beta}\right) - \ln(\delta)}{\sqrt{\alpha^2 \times v^2_{\frac{self,A}{other,B}} + v^2_{\frac{\beta}{1-\beta}}}}\right)$; $Pr([\text{A} \succ \text{B} \times 1/2]) = \Phi\left(\frac{\alpha \times \ln\left(\frac{A}{B}\right) - \ln\left(\frac{1}{2}\right) - \ln\left(\frac{1}{\mu_{1-\alpha}}\right)}{\alpha \times v_{\frac{self,A}{other,B}}}\right)$	-15,103.79	-10,519.21
Preferences Noise $v_{\frac{\beta}{1-\beta}} = 0$ Combined	$Pr([\text{self} \succ \text{other}]) = \Phi\left(\frac{\alpha \times \ln\left(\frac{self}{other}\right) - \ln\left(\frac{\beta}{1-\beta}\right) - \ln\left(\frac{1}{\mu_{1-\alpha}}\right)}{\alpha \times v_{\frac{self,A}{other,B}}}\right)$; $Pr([\text{A} \succ \text{B} \times 1/2]) = \Phi\left(\frac{\alpha \times \ln\left(\frac{A}{B}\right) - \ln\left(\frac{1}{2}\right) - \ln\left(\frac{1}{\mu_{1-\alpha}}\right)}{\alpha \times v_{\frac{self,A}{other,B}}}\right)$	-15,640.32	-11,853.78

(a) Models Altruism and Number Comparison



(b) $ELPD_{WAIC,A}$ Altruism



(c) $ELPD_{WAIC,NC}$ Number Comparison

Figure A8. Model Comparison Combined Models (a) Altruism Choices ($ELPD_{WAIC,A}$) and (b) Number Comparison ($ELPD_{WAIC,NC}$).

A.3.2 Combined Estimation.

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 A2 and A4 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 A11, 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 columns thus do not provide evidence for a learned treatment effect and instead, point towards some attenuation over time in the number comparison data.

Table A11. Treatment effect and learning regression

	(1)	(2)	(3)	(4)
Treatment Group	-0.01430**	-0.01430	-0.04916***	-0.05084***
	(0.00603)	(0.02342)	(0.00362)	(0.00690)
Ratio $\frac{\text{self}}{\text{other}}$ / $\frac{\text{A}}{\text{B}}$	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)
N	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
R^2	0.430	0.517	0.791	0.679

Note: Linear Probability Model. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky & 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 $\bar{\text{self}}_T = 0.393$, $\bar{\text{self}}_B = 0.367$ ($p = 0.6356$). However, in the 12 practice trials[33] the Treatment group decides significantly more often for the other person with an average of $\bar{\text{self}}_T = 0.364$, $\bar{\text{self}}_B = 0.414$ ($p < 0.01$). 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 explanation: Participants quickly understand how the task works, i.e., "less-for-me" vs

33 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.

"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, this data suggests 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 then 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; Özdemir et al., 2010; Schwappach & Strasmann, 2006). My 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.

Table A12. Inconsistency regression on trial level

	(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
N	14400	14400	12000	12000
R^2	0.131	0.206	0.066	0.309

Note: Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky & Tipton, 2018)) in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

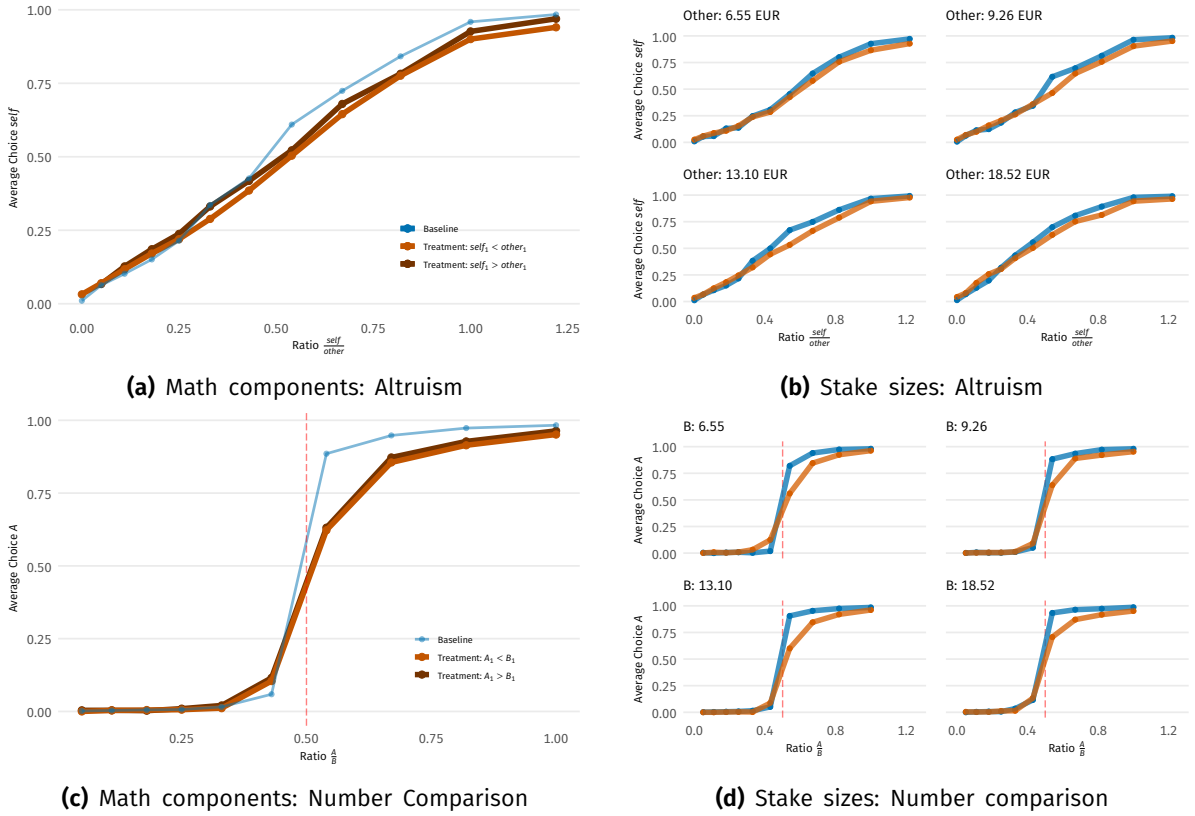
Table A12 performs a linear regression on a dataset with 300 (subject) × 48 (trial groups, see Figure A3) = 14,400 respectively 200 × 40 = 12,000 observations based on the Altruism and Number Comparison data. Each row of this dataset contains the

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 A12, 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 A17). 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 $self_1 < other_1$.

Panel (a) of Figure A9 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 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).



(a) Math components: Altruism

(b) Stake sizes: Altruism

(c) Math components: Number Comparison

(d) Stake sizes: Number comparison

Figure A9. 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 $\frac{\text{self}}{\text{other}}$, $\frac{\text{A}}{\text{B}}$, for Baseline and Treatment whereas the latter is divided into cases where the first math component $\text{self}_1 > \text{other}_1$ and $\text{self}_1 < \text{other}_1$, with $\text{self} = \text{self}_1 + \text{self}_2$ and $\text{other} = \text{other}_1 + \text{other}_2$. Panels (b) and (d) plot average choices separately for the different base values of other and B.

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 A9 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 A13 performs a regression analysis akin to the linear probability models in Tables A2 and A4 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.

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 & Athey, 2018). Causal Forests

Table A13. Stake-size and choice regression

	(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)
N	72000	60000
Data	Altruism	Number Comp.
Clustered Standard Errors	Yes	Yes
Unique Obs	300	300
R^2	0.008	0.002

Note: Linear Probability Model. Clustered standard errors (participant-level, "bias-reduced linearization" (Pustejovsky & Tipton, 2018)) in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

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).[34] The main advantage over classical techniques (i.e., interaction terms 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.

Figure A10 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. The key take-away is that (i) the variation in most personal characteristics does 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

34 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.

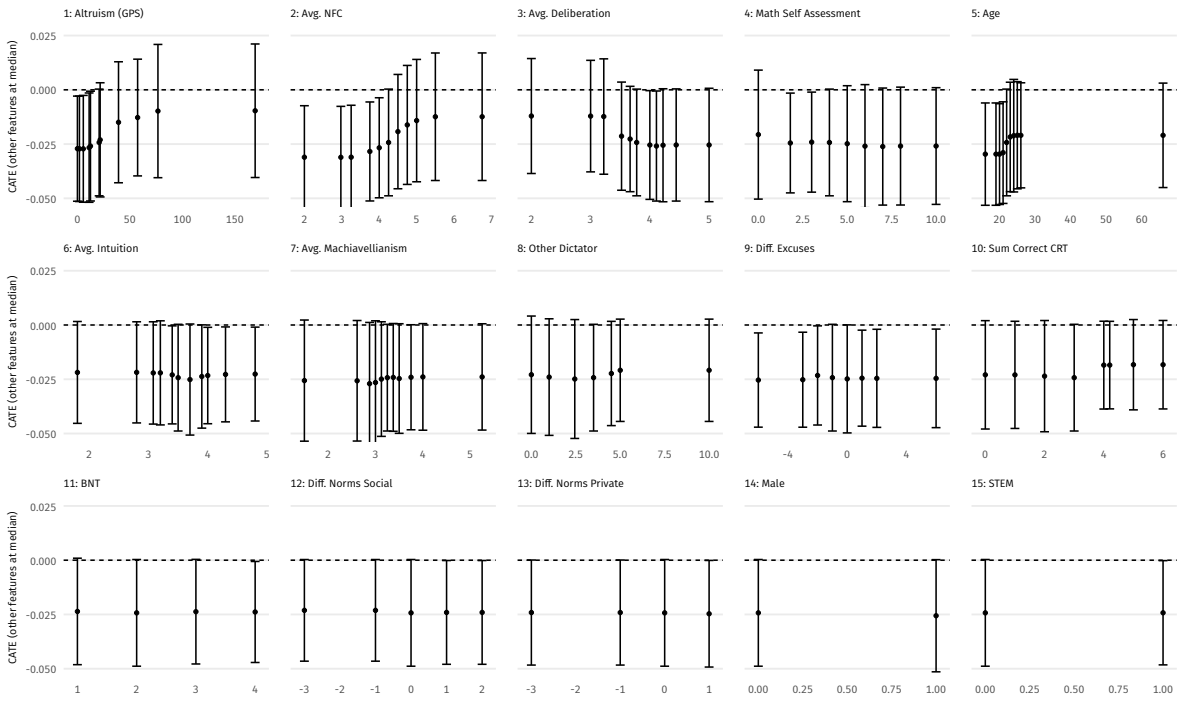


Figure A10. Heterogeneous Treatment Effects This plot shows the estimated CATE values for each personal characteristic (sorted by their importance), estimated at each decile of the feature distribution, 95 % confidence intervals.

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 the present dataset is not large enough to detect minuscule differences in treatment heterogeneity.

A.3.5 Correlation between Metacognition, RT and Choices. Table A14 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 a stronger association between choices, inconsistencies and RT with metacognitive measures in the number comparison domain compared to the altruism choices.

Table A14. Correlation metacognition altruistic choice and number comparison

	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

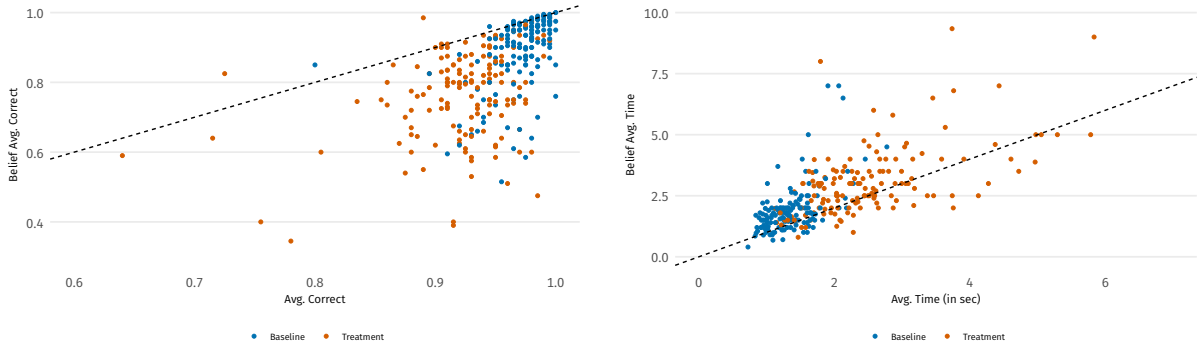
Note: p-values from pairwise rank-correlation tests ($n = 300$). * p < 0.1, ** p < 0.05, *** p < 0.01.

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 $|\Delta$ 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 A14, 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 ($\rho = -0.231$). 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 ($\rho = -0.151$).

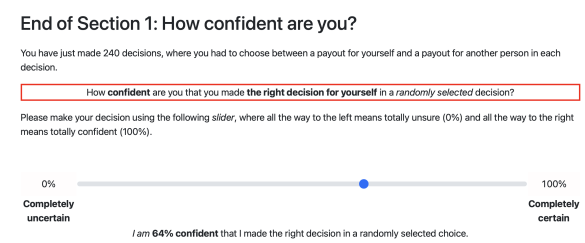


(a) Beliefs (Avg.) Correctly Solved Tasks

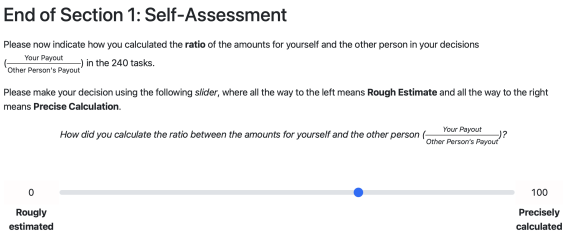
(b) Beliefs (Avg.) Time Spent

Figure A11. Beliefs Number Comparison This plot depicts the number comparison belief data, in correctly solved tasks (a) and the average time spent (b) against either objective counterpart.

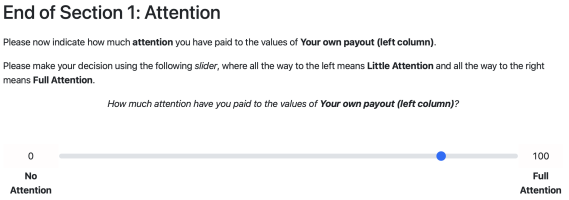
A.4 Experimental Screenshots



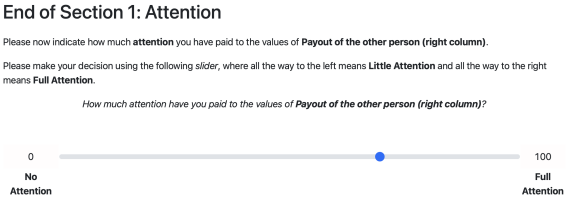
(a) Confidence



(b) Precision



(c) Attention self



(d) Attention other

Figure A12. Screenshots: Metacognition Altruistic Choice

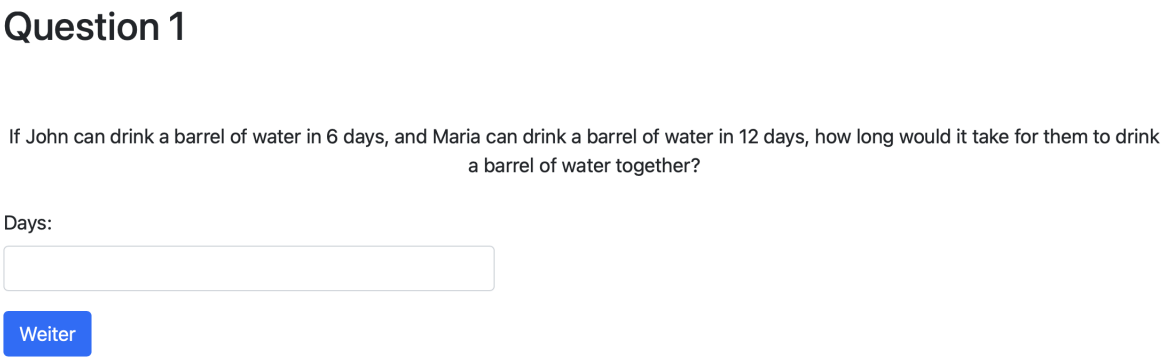
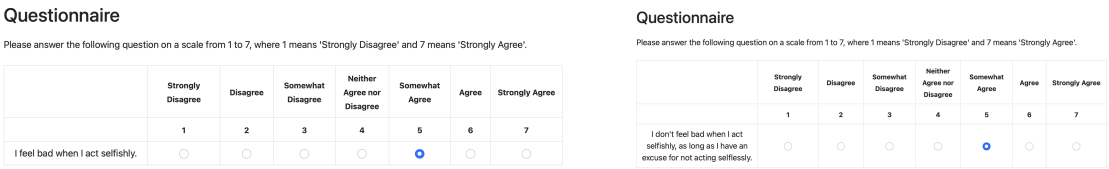


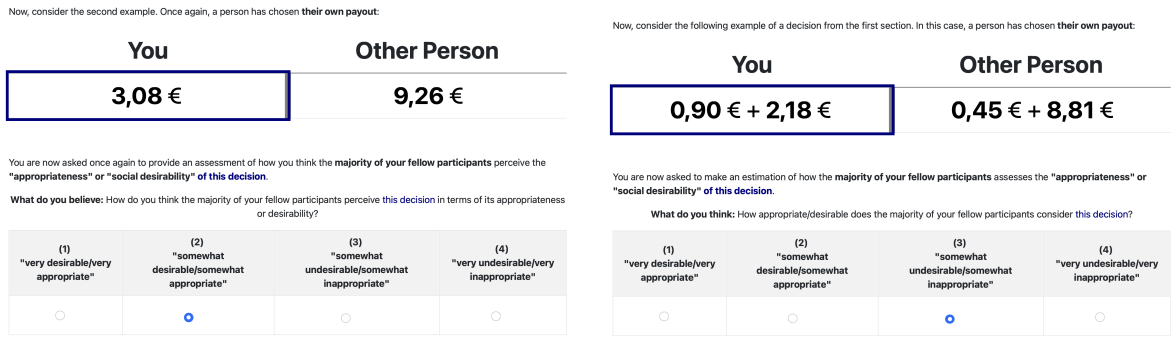
Figure A13. Screenshot: CRT4 Question



(a) No Excuses

(b) Excuses (reverse formulated)

Figure A14. 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 A15. Screenshots: Social Norms **(a)** Baseline **(b)** Treatment. Both variants are shown to all participants, order in which questions appear is randomized

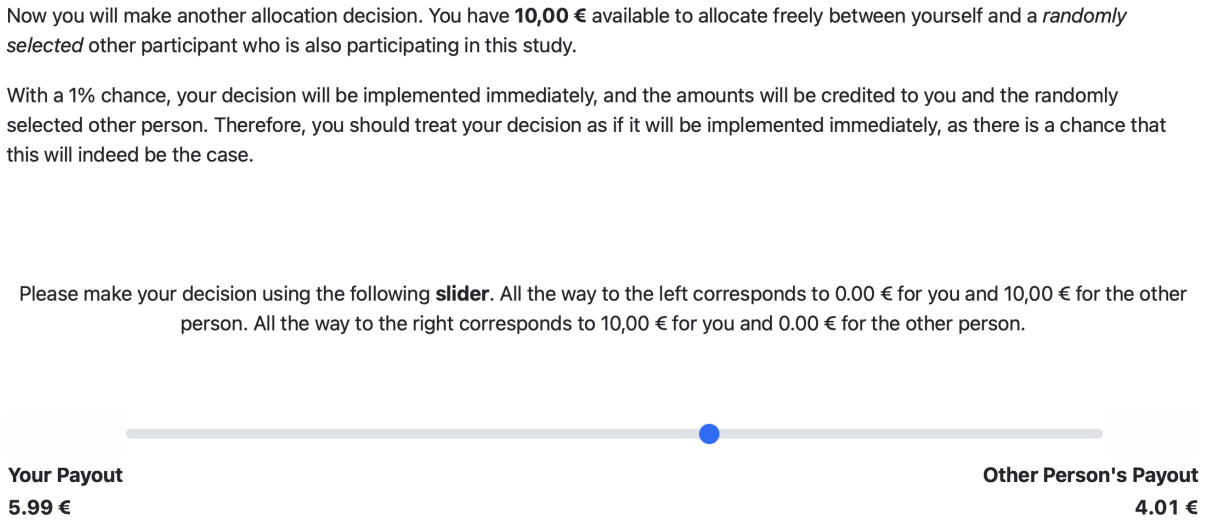


Figure A16. Screenshot: Dictator Game

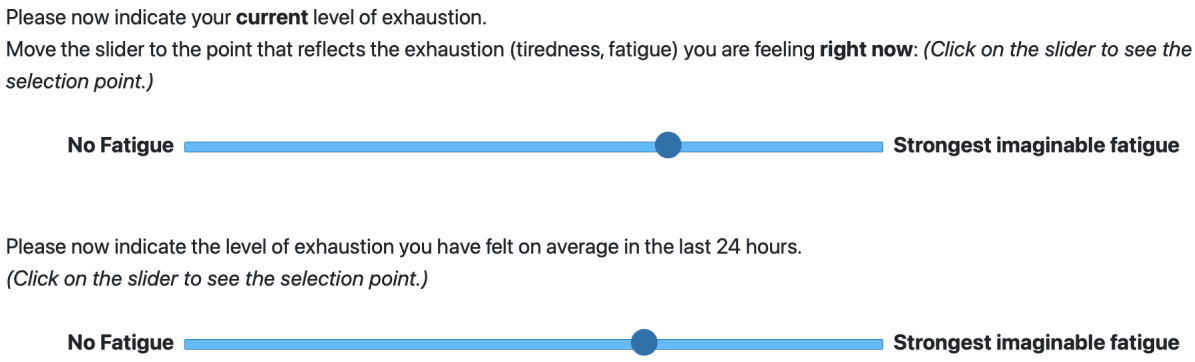


Figure A17. Screenshot: Fatigue Visual Analog Scales

References

Alós-Ferrer, C., & Garagnani, M. (2022). Strength of preference and decisions under risk. *Journal of Risk and Uncertainty*, *64*(3), 309–329. https://doi.org/10.1007/s11166-022-09381-0. [35, 41]

Alós-Ferrer, C., Garagnani, M., & Hügelschäfer, S. (2016). Cognitive Reflection, Decision Biases, and Response Times. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01402. [37]

Amador-Hidalgo, L., Brañas-Garza, P., Espín, A. M., García-Muñoz, T., & Hernández-Román, A. (2021). Cognitive abilities and risk-taking: Errors, not preferences. *European Economic Review*, *134*, 103694. https://doi.org/10.1016/j.euroecorev.2021.103694. [43]

Amelio, A. (2022). Cognitive Uncertainty and Overconfidence. *ECONtribute Discussion Paper No. 173*. [39]

Andreoni, J. (1989). Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence. *Journal of Political Economy*, *97*(6), 1447–1458. https://doi.org/10.1086/261662. [7]

Andreoni, J., & Miller, J. (2002). Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica*, *70*(2), 737–753. https://doi.org/10.1111/1468-0262.00302. [2, 7]

Assenza, T., Cardaci, A., & Delli Gatti, D. (2019). Perceived Wealth, Cognitive Sophistication and Behavioral Inattention. *CESifo Working Paper No. 7992*. https://doi.org/10.2139/ssrn.3507263. [13, 34]

Augenblick, N., Lazarus, E., & Thaler, M. (2022). Overinference from Weak Signals and Underinference from Strong Signals. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4315007. [13, 34]

Bao, T., & Pei, J. (2024). Cognitive Uncertainty, GPT, and Contribution in Public Goods Game. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4525626. [3, 42]

Barretto-García, M., De Hollander, G., Grueschow, M., Polanía, R., Woodford, M., & Ruff, C. C. (2023). Individual risk attitudes arise from noise in neurocognitive magnitude representations. *Nature Human Behaviour*, *7*(9), 1551–1567. https://doi.org/10.1038/s41562-023-01643-4. [9, 40, 41]

Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oka, P., Oprescu, M., & Syrgkanis, V. (2019). EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. [67]

Bech, M., Kjaer, T., & 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. https://doi.org/10.1002/hec.1587. [64]

Beißert, H., Köhler, M., Rempel, M., & Beierlein, C. (2014). Eine deutschsprachige Kurzskala zur Messung des Konstrukts Need for Cognition. *GESEIS Working Papers*, *32*. [18]

Bellemare, C., Kröger, S., & 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. https://doi.org/10.1016/j.jebo.2011.01.019. [2, 5, 6]

Bernheim, B. D., & Stark, O. (1988). Altruism within the Family Reconsidered: Do Nice Guys Finish Last? *The American Economic Review*, *78*(5), 1034–1045. [6]

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. [18]

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P. A., Horsfall, P., & Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, *20*, 28:1–28:6. [22]

Bland, J. R. (2023). Bayesian Model Selection and Prior Calibration for Structural Models in Economic Experiments: Some Guidance for the Practitioner. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4334267. [21]

Bolton, G. E., & Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, *90*(1), 166–193. https://doi.org/10.1257/aer.90.1.166. [2]

Bruhin, A., Fehr, E., & 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. [2, 5, 20]

Brus, J., Aebersold, H., Grueschow, M., & Polania, R. (2021). Sources of confidence in value-based choice. *Nature Communications*, *12*(1), 7337. https://doi.org/10.1038/s41467-021-27618-5. [43]

Burks, S. V., Carpenter, J. P., Goette, L., & Rustichini, A. (2009). Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences*, *106*(19), 7745–7750. https://doi.org/10.1073/pnas.0812360106. [42]

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116. [18]

Cantlon, J. F., Libertus, M. E., Pinel, P., Dehaene, S., Brannon, E. M., & Pelphrey, K. A. (2009). The Neural Development of an Abstract Concept of Number. *Journal of Cognitive Neuroscience*, *21*(11), 2217–2229. https://doi.org/10.1162/jocn.2008.21159. [8]

Cappelen, A. W., Nielsen, U. H., Tungodden, B., Tyran, J.-R., & Wengström, E. (2016). Fairness is intuitive. *Experimental Economics*, *19*(4), 727–740. https://doi.org/10.1007/s10683-015-9463-y. [5, 35]

Carpenter, J., & Robbett, A. (2024). Measuring socially appropriate social preferences. *Games and Economic Behavior*, *147*, 517–532. https://doi.org/10.1016/j.geb.2024.08.003. [2, 5, 6, 20]

Chapman, J., Dean, M., Ortoleva, P., Snowberg, E., & Camerer, C. (2023). Econographics. *Journal of Political Economy Microeconomics*, *1*(1), 115–161. https://doi.org/10.1086/723044. [42]

Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. [2, 7]

Chen, C.-C., Chiu, I.-M., Smith, J., & Yamada, T. (2013). Too smart to be selfish? Measures of cognitive ability, social preferences, and consistency. *Journal of Economic Behavior & Organization*, *90*, 112–122. https://doi.org/10.1016/j.jebo.2013.03.032. [42]

Chew, S. H., Miao, B., Shen, Q., & Zhong, S. (2022). Multiple-switching behavior in choice-list elicitation of risk preference. *Journal of Economic Theory*, *204*, 105510. https://doi.org/10.1016/j.jet.2022.105510. [13, 34]

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring Risk Literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, *7*(1), 25–47. https://doi.org/10.1017/S1930297500001819. [18]

Cooper, D. J., & Kagel, J. H. (2016). Other-Regarding Preferences A Selective Survey of Experimental Results. In J. H. Kagel & A. E. Roth (Eds.), *The Handbook of Experimental Economics, Volume Two*. Princeton University Press. https://doi.org/10.1515/9781400883172-005. [2]

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. [18]

Dehaene, S. (1993). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Sensorimotor foundations of higher cognition* (pp. 527–574). [8]

Dehaene, S. (2011). *The Number Sense: How the Mind Creates Mathematics, Revised and Updated Edition* (Updated Edition). Oxford University Press. [7]

Dehaene, S., & 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. https://doi.org/10.1080/02724980244000044. [15]

Diester, I., & Nieder, A. (2007). Semantic Associations between Signs and Numerical Categories in the Prefrontal Cortex (S. Dehaene, Ed.). *PLoS Biology*, *5*(11), e294. https://doi.org/10.1371/journal.pbio.0050294. [7]

Doya, K., Ishii, S., Pouget, A., Rao, R. P. N., Sejnowski, T. J., & Poggio, T. A. (Eds.). (2006). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. MIT Press. [7]

Echeverry, D., Figueroa, M. C., & Polanıa-Reyes, S. (2023). Structural Identification of Social Preferences: Heterogeneity Matters for Incentives. *University de Nvarra Working Papers*, (2). [5]

Enke, B. (2024). The Cognitive Turn in Behavioral Economics. [2, 5, 42]

Enke, B., & Graeber, T. (2023). Cognitive Uncertainty. *The Quarterly Journal of Economics*, *138*(4), 2021–2067. [2, 38, 39, 42]

Enke, B., Graeber, T., & Oprea, R. (2023). Complexity and Hyperbolic Discounting. [2, 4, 15, 42]

Enke, B., Graeber, T., Oprea, R., & Yang, J. (2024). Behavioral Attenuation. [3, 42]

Exley, C., & Kessler, J. (2024). Motivated Errors. *American Economic Review*, *114*(4), 961–987. https://doi.org/10.3386/w26595. [18]

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global Evidence on Economic Preferences. *The Quarterly Journal of Economics*, *133*(4), 1645–1692. https://doi.org/10.1093/qje/qjy013. [42]

Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2023). The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences. *Management Science*, *69*(4), 1935–1950. https://doi.org/10.1287/mnsc.2022.4455. [18]

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, *54*(2), 293–315. https://doi.org/10.1016/j.geb.2005.03.001. [6]

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. [2, 6]

Fehr, E., & Charness, G. (2023). Social Preferences: Fundamental Characteristics and Economic Consequences. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4464745. [2]

Fehr, E., & 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* (pp. 615–691, Vol. 1). Elsevier. https://doi.org/10.1016/S1574-0714(06)01008-6. [2]

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. https://doi.org/10.1016/j.tics.2004.05.002. [7]

Fisman, R., Kariv, S., & Markovits, D. (2007). Individual Preferences for Giving. *American Economic Review*, *97*(5), 1858–1876. [2, 5, 7]

Fleming, S. M. (2024). Metacognition and Confidence: A Review and Synthesis. *Annual Review of Psychology*, (75), 241–268. [43]

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732. [13, 17]

Frydman, C., & Jin, L. J. (2022). Efficient Coding and Risky Choice. *The Quarterly Journal of Economics*, *137*(1), 161–213. https://doi.org/10.1093/qje/qjab031. [2, 4, 13, 17, 32, 41]

Frydman, C., & Nunnari, S. (2023). Coordination with Cognitive Noise. [13, 17]

Gabaix, X. (2019). Behavioral inattention. In *Handbook of Behavioral Economics: Applications and Foundations 1* (pp. 261–343, Vol. 2). Elsevier. [9]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2021). Bayesian Data Analysis. (Third Edition). [21]

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543. https://doi.org/10.3102/1076998615606113. [22]

Graf, C., Vetschera, R., & Zhang, Y. (2013). Parameters of social preference functions: Measurement and external validity. *Theory and Decision*, *74*(3), 357–382. https://doi.org/10.1007/s11238-012-9312-9. [5]

Hauge, K. E., Brekke, K. A., Johansson, O., Johansson, O., & Svedsäter, H. (2009). Are Social Preferences Skin Deep? Dictators under Cognitive Load. *University of Gothenburg Working Papers in Economics*, (No. 371). [42]

Holzmeister, F., & Stefan, M. (2021). The risk elicitation puzzle revisited: Across-methods (in)consistency? *Experimental Economics*, *24*(2), 593–616. https://doi.org/10.1007/s10683-020-09674-8. [42]

Hossain, T., & Okui, R. (2013). The Binarized Scoring Rule. *The Review of Economic Studies*, *80*(3), 984–1001. https://doi.org/10.1093/restud/rdt006. [17]

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron*, *87*(2), 451–462. https://doi.org/10.1016/j.neuron.2015.06.031. [6]

Khaw, M. W., Li, Z., & Woodford, M. (2021). Cognitive Imprecision and Small-Stakes Risk Aversion. *The Review of Economic Studies*, *88*(4), 1979–2013. [2, 4, 6, 7, 14, 15, 40, 41]

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, *84*(6), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015. [43]

Klockmann, V., Von Schenk, A., & Villeval, M. C. (2022). Artificial intelligence, ethics, and intergenerational responsibility. *Journal of Economic Behavior & Organization*, *203*, 284–317. https://doi.org/10.1016/j.jebo.2022.09.010. [2, 5, 20]

König-Kersting, C. (2024). On the robustness of social norm elicitation. *Journal of the Economic Science Association*. https://doi.org/10.1007/s40881-024-00178-2. [18]

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, *6*(1), 7455. https://doi.org/10.1038/ncomms8455. [35]

Krupka, E. L., & 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. https://doi.org/10.1111/jeea.12006. [18]

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press. [21]

Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, *4*(33), 1143. https://doi.org/10.21105/joss.01143. [60]

Laurent, G., & Vanhuele, M. (2023). How Do Consumers Read and Encode a Price? (A. Kirmani, J. Cotte, & M. Thomas, Eds.). *Journal of Consumer Research*, *50*(3), 510–532. https://doi.org/10.1093/jcr/ucad005. [15]

Lepper, M. (2024). Excuse-Based Procrastination. *mimeo*. [71]

Levine, D. K. (1998). Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics*, *1*(3), 593–622. https://doi.org/10.1006/redy.1998.0023. [2, 6, 7]

Longo, M. R., & Lourenco, S. F. (2007). Spatial attention and the mental number line: Evidence for characteristic biases and compression. *Neuropsychologia*, *45*(7), 1400–1407. https://doi.org/10.1016/j.neuropsychologia.2006.11.002. [8]

Luce, R. D. (1991). *Response times: Their role in inferring elementary mental organization* (1. issued as paperback). Univ. Press. [34]

McFadden, D. (1981). Econometric Models for Probabilistic Choice. In C. Manski & D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press. [5]

Merkel, A. L., & Lohse, J. (2019). Is fairness intuitive? An experiment accounting for subjective utility differences under time pressure. *Experimental Economics*, *22*(1), 24–50. https://doi.org/10.1007/s10683-018-9566-3. [35]

Moyer, R. S., & Landauer, T. K. (1967). Time required for Judgements of Numerical Inequality. *Nature*, *215*(5109), 1519–1520. https://doi.org/10.1038/2151519a0. [35]

Natenzon, P. (2019). Random Choice and Learning. *Journal of Political Economy*, *127*(1), 419–457. https://doi.org/10.1086/700762. [22]

Nieder, A., & Dehaene, S. (2009). Representation of Number in the Brain. *Annual Review of Neuroscience*, *32*(1), 185–208. https://doi.org/10.1146/annurev.neuro.051508.135550. [7]

Nunnari, S., & Pozzi, M. (2024). Meta-Analysis of Distributional Preferences Estimates. [2, 5, 20]

Olschewski, S., Rieskamp, J., & Hertwig, R. (2023). The link between cognitive abilities and risk preference depends on measurement. *Scientific Reports*, *13*(1), 21151. https://doi.org/10.1038/s41598-023-47844-9. [43]

Olschewski, S., Rieskamp, J., & Scheibehenne, B. (2018). Taxing cognitive capacities reduces choice consistency rather than preference: A model-based test. *Journal of Experimental Psychology: General*, *147*(4), 462–484. https://doi.org/10.1037/xge0000403. [42]

Olschewski, S., & Scheibehenne, B. (2024). What's in a sample? Epistemic uncertainty and metacognitive awareness in risk taking. *Cognitive Psychology*, *149*, 101642. https://doi.org/10.1016/j.cogpsych.2024.101642. [38]

Oprea, R. (2024). Decisions Under Risk are Decisions Under Complexity. *American Economic Review*. [2, 13, 15, 34, 38]

Oprea, R., & Vieider, F. (2024). Minding the Gap: On the Origins of Probability Weighting and the Description-Experience Gap. [5, 22, 40]

Özdemir, S., Mohamed, A. F., Johnson, F. R., & Hauber, A. B. (2010). Who pays attention in stated-choice surveys? *Health Economics*, *19*(1), 111–118. https://doi.org/10.1002/hec.1452. [64]

Phan, D., Pradhan, N., & Jankowiak, M. (2019). Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv preprint arXiv:1912.11554*. [22]

Pins, D., & 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. https://doi.org/10.3758/BF03206815. [35]

Poggi, L. (2021). Learning dynamics in optimal decision making. *mimeo*. [41]

Polanía, R., Woodford, M., & Ruff, C. C. (2019). Efficient coding of subjective value. *Nature Neuroscience*, *22*(1), 134–142. https://doi.org/10.1038/s41593-018-0292-0. [2, 8, 41]

Ponti, G., & Rodriguez-Lara, I. (2015). Social preferences and cognitive reflection: Evidence from a dictator game experiment. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00146. [42]

Prat-Carrabin, A., & Gershman, S. J. (2024). Bayes vs. Weber: How to break a law of psychophysics. [41]

Prat-Carrabin, A., & Woodford, M. (2022). Efficient coding of numbers explains decision bias and noise. *Nature Human Behaviour*, *6*(8), 1142–1152. https://doi.org/10.1038/s41562-022-01352-4. [8]

Pustejovsky, J. E., & 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. https://doi.org/10.1080/07350015.2016.1247004. [37, 55, 56, 63, 64, 67]

Radbruch, L., Sabatowski, R., Elsner, F., Everts, J., Mendoza, T., & Cleeland, C. (2003). Validation of the German Version of the Brief Fatigue Inventory. *Journal of Pain and Symptom Management*, *25*(5), 449–458. https://doi.org/10.1016/S0885-3924(03)00073-3. [18, 65]

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430. https://doi.org/10.1038/nature11467. [5, 35]

Rand, D. G., & Kraft-Todd, G. T. (2014). Reflection does not undermine self-interested prosociality. *Frontiers in Behavioral Neuroscience*, *8*. https://doi.org/10.3389/fnbeh.2014.00300. [35]

Schlag, K., & Van Der Weele, J. (2013). Eliciting Probabilities, Means, Medians, Variances and Covariances without Assuming Risk Neutrality. *Theoretical Economics Letters*, *3*(1), 38–42. https://doi.org/10.4236/tel.2013.31006. [17]

Schunk, D., & Betsch, C. (2006). Explaining heterogeneity in utility functions by individual differences in decision modes. *Journal of Economic Psychology*, *27*(3), 386–401. https://doi.org/10.1016/j.joep.2005.08.003. [18]

Schwappach, D. L., & 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. https://doi.org/10.1016/j.jhealeco.2005.08.002. [64]

Stango, V., & 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. https://doi.org/10.1093/restud/rdac055. [42]

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*(4), 578–585. https://doi.org/10.1038/nn1669. [16]

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729. [17]

Van Leeuwen, B., & Alger, I. (2024). Estimating Social Preferences and Kantian Morality in Strategic Interactions. *Journal of Political Economy Microeconomics*, *2*(4), 665–706. https://doi.org/10.1086/732125. [5]

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4. [24]

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-BA1221. [22]

Vieider, F. (2024a). Bayesian Estimation of Decision Models. [4, 21]

Vieider, F. (2024b). Decisions Under Uncertainty as Bayesian Inference on Choice Options. *Management Science*. https://doi.org/10.1287/mnsc.2023.00265. [2, 3, 6–8, 10, 36, 40]

Vieider, F. M. (2023). Cognitive Foundations of Delay-Discounting. [4]

Von Schenk, A., Klockmann, V., & Köbis, N. (2023). Social Preferences Toward Humans and Machines: A Systematic Experiment on the Role of Machine Payoffs. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916231194949. [20]

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839. [66]

Watanabe, S. (2013). A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research*, *14*, 867–897. [24]

Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, *114*(38), 10244–10249. https://doi.org/10.1073/pnas.1619153114. [8]

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604. https://doi.org/10.1038/nn0602-858. [16]

Woodford, M. (2012). Prospect Theory as Efficient Perceptual Distortion. *American Economic Review*, *102*(3), 41–46. https://doi.org/10.1257/aer.102.3.41. [2]

Woodford, M. (2020). Modeling Imprecision in Perception, Valuation, and Choice. *Annual Review of Economics*, *12*(1), 579–601. [2, 4]

Xie, Y., & Carlin, B. P. (2006). Measures of Bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, *136*(10), 3458–3477. https://doi.org/10.1016/j.jspi.2005.04.003. [29]