Measuring Hearts and Minds: A Validated Survey Module on Inequality Aversion and Altruism

Thomas F. Epper[§] and Ivan Mitrouchev[†]

[§]CNRS, IESEG School of Management, Univ. Lille, UMR 9221 - LEM - Lille Economie Management, F-59000 Lille, France. thomas.epper@cnrs.fr. ORCID n°0000-0002-0826-4997.
[†]Univ. Grenoble Alpes, INRAE, CNRS, Grenoble INP, GAEL, 38000 Grenoble, France. ivan.mitrouchev@inrae.fr. ORCID n°0000-0002-8960-4550.

January 7, 2025

Abstract

Social preferences, including trust, altruism, and reciprocity, are widely studied in behavioral economics, with validated survey modules available to measure these traits. However, despite growing interest in inequality aversion—defined as an individual's dislike of disparities in outcomes—there is no dedicated and validated module to assess this specific social preference. Moreover, the relationship between inequality aversion and altruism is not always explicitly addressed in existing frameworks. To bridge these gaps, we introduce a novel survey module that captures general attitudes toward altruism while integrating measures of inequality aversion, reflecting the inherent connection between these two factors. This module was developed and validated through an experimental study with a representative U.S. population sample (n = 502). Our results demonstrate that the proposed module effectively captures variations in both inequality aversion and altruism, with consistent reliability across individual heterogeneity. This new tool offers researchers a standardized and generalizable approach for measuring inequality aversion and altruism, paving the way for future studies in these areas across diverse contexts.

Keywords. *inequality — altruism — redistribution — social preferences*

JEL codes. D63, D91

Statements and declarations. Thomas F. Epper acknowledges funding from the Métropole Européenne de Lille (MEL). Ivan Mitrouchev acknowledges the financial support of the FAST project (Facilitate public Action to exist from peSTicides) conducted by the Agence Nationale de la Recherche (ANR), reference 20-PCPA-0005. There are no competing interests.

1 Introduction

Incentivized experiments have traditionally been the preferred method for measuring social preferences, but these can be resource-intensive, requiring substantial time and financial input. Experiments typically provide real incentives, which can help mitigate issues related to inattention and hypothetical bias in survey responses. However, the costs associated with incentivized experiments can make them challenging to implement at scale, particularly in field settings or broad population studies. The recent literature has largely contributed to validated modules on risk aversion, time discounting, trust, altruism, positive and negative reciprocity (Falk et al., 2018, 2023). While inequality aversion—defined as individuals' sensitivity to unequal distributions of resources—has been a cornerstone of theoretical models (Fehr and Schmidt, 1999) and extensively examined in the context of preferences for redistribution (Fong, 2001; Alesina and Angeletos, 2005; Alesina and Giuliano, 2011), it, however, remains underexplored in terms of standardized measurement tools. In particular, most studies focus on redistribution in specific policy settings (Corneo and Grüner, 2002; Guillaud, 2013; Hvidberg et al., 2023), in relation to welfare (Decancq et al., 2017, 2019; Fleurbaey and Zuber, 2024), or within macroeconomic contexts (Piketty and Saez, 2003, 2014), rather than providing generalizable tools for capturing inequality aversion at the individual level.[1] This gap underscores the need for a validated survey module that integrates inequality aversion into a broader framework of social preferences, enabling researchers to study its role in diverse contexts and populations.

We aim to address this gap by proposing a survey module that jointly measures inequality aversion and altruism. Our goal is to develop a tool that retains the predictive accuracy of incentivized experiments while being parsimonious and feasible for broad application. Our approach is based on an online experimental study with a U.S. general population sample (n = 502). Using this sample, we calibrate and validate the survey module, selecting items based on their ability to predict choices in an incentivized preference elicitation task. For survey item selection, we implemented a rigorous procedure, testing a wide range of candidate items and identifying those with the highest predictive accuracy. We employ machine learning techniques (a gradient boosting algorithm) to predict preference types and inequality aversion. We then evaluate the predictive ability of the large set of survey items using SHapley Additive exPlanations (SHAP values).[2] We eventually identify a restricted set of survey items that, when appropriately weighted, explain a reasonably large proportion of behavioral type and parameter variation we observe in the sample and its subsets. Although our proposed module was validated using a general population sample (with various subsets of the data used for training and validation), we anticipate that it will serve as a useful measure of inequality aversion across various populations and cultural contexts. While some predictive power may be sacrificed compared to incentivized experiments, this trade-off allows for a more accessible and cost-effective measurement tool. Additionally, the module's transparency and the methodology used in item selection allow researchers to easily adapt the tool to

[1] For a literature review on preferences for redistribution, see also Mengel and Weidenholzer (2023). For an introduction to the concept of economic inequality, see Cowell (2011).

[2] SHAP provide a method for interpreting machine learning models by attributing the contribution of each feature (variable) to a model's predictions. This approach is based on the Shapley value concept from cooperative game theory (Shapley, 1953).

meet specific research needs.

The remainder of the paper is structured as follows. Section 2 describes the research design, with the characterization of the sample, the description of the preference elicitation method used in the experiment, the tested survey items, as well as the hypothetical choices and real-world behavior. Section 3 presents the method and results to infer preference types and parameter from the incentivized preference elicitation task. In Section 4 we provide analyses of the survey responses, their association with the preference types, and their correlation with inequality aversion parameters. Section 5 develops a predictive model that uses a concise subset of survey responses to predict both type assignment and differences in inequality aversion. We evaluate this model's performance in predicting self-reported real-world social actions and compare its predictive accuracy to that of the more resource-intensive incentivized measures. Finally, Section 6 concludes with an overview of potential applications and future research directions.

2 Research Design

This section provides an overview of the sample and study design. In Section 2.1, we describe the sample's characteristics and evaluate its representativeness of the U.S. adult population. Section 2.2 outlines the incentivized preference elicitation task, which serves as the core of our analysis. Section 2.3 details the comprehensive set of survey items included in the study. Finally, Section 2.4 introduces a series of hypothetical questions and real-world behavior.

2.1 Setup and Sample

The study was conducted online using a representative sample of the U.S. adult population. A total of 536 participants, recruited *via* Prolific in autumn 2024, completed the study. The online sessions lasted approximately 40 minutes on average. Participants received a fixed completion fee of £4, along with a variable bonus payment based on one randomly selected decision from the preference elicitation task.[3] The variable bonus payments ranged from £2.70 to £7.05. Our analysis focuses on participants who successfully passed three attention checks administered throughout the study. Details of the attention checks are provided in Appendix A.1. Of those completing the study, a high proportion of 93.7% met this criterion, yielding a final dataset of 502 participants. As shown in Appendix A.2, this restricted sample remains broadly representative of the U.S. adult population across three key stratification criteria: age group, gender and ethnicity.

2.2 Preference Elicitation

To measure respondents' social preferences, we employ a series of 20 money allocation tasks. The task design is adapted from Fehr et al. (2024) and Epper et al. (2024).[4] Calibration follows Fehr et al. (2023), who elicited distributional preferences from a Swiss

[3]The British Pound (£) is the default currency used by Prolific, regardless of the participant's country of residence.

[4]Fehr et al. (2024) elicit social preferences of a Swiss broad population sample using a larger set of (64) choice situations. Epper et al. (2024) elicit social preferences of a Danish broad population sample using 11 (instead of 7) choice options per situation and a slightly different configuration.

representative sample (referring to their 2020 wave). The primary modification in our study is a shift in the payoffs, with 100 ECU (experimental currency units) equivalent to £1. The 20 choice situations are listed in Table 1.

Table 1: Choice situations

j	y^s	x^s	y^o	x^o	domain	MCR	α_{crit}	β_{crit}
1	270.0	630.0	450.0	450.0	mixed	-Inf	0.00	0.00
2	300.0	600.0	480.0	420.0	mixed	5.00	-0.83	0.83
3	330.0	570.0	510.0	390.0	mixed	2.00	-0.67	0.67
4	360.0	540.0	540.0	360.0	mixed	1.00	-0.50	0.50
5	390.0	510.0	570.0	330.0	mixed	0.50	-0.33	0.33
6	420.0	480.0	600.0	300.0	mixed	0.20	-0.17	0.17
7	450.0	450.0	630.0	270.0	mixed	0.00	-0.00	0.00
8	420.0	480.0	300.0	600.0	mixed	-0.20	0.25	-0.25
9	390.0	510.0	330.0	570.0	mixed	-0.50	1.00	-1.00
10	360.0	540.0	360.0	540.0	mixed	-1.00	Inf	-Inf
11	330.0	570.0	390.0	510.0	mixed	-2.00	-2.00	2.00
12	300.0	600.0	420.0	480.0	mixed	-5.00	-1.25	1.25
13	420.0	460.2	480.0	679.8	behind	-0.20	0.25	
14	480.0	570.0	420.0	150.0	ahead	0.33		0.25
15	420.0	480.0	480.0	660.0	behind	-0.33	0.50	
16	480.0	660.0	420.0	240.0	ahead	1.00		0.50
17	420.0	492.0	480.0	648.0	behind	-0.43	0.75	
18	480.0	705.0	420.0	345.0	ahead	3.00		0.75
19	420.0	430.8	480.0	711.0	behind	-0.05	0.05	
20	480.0	498.0	420.0	78.0	ahead	0.05		0.05

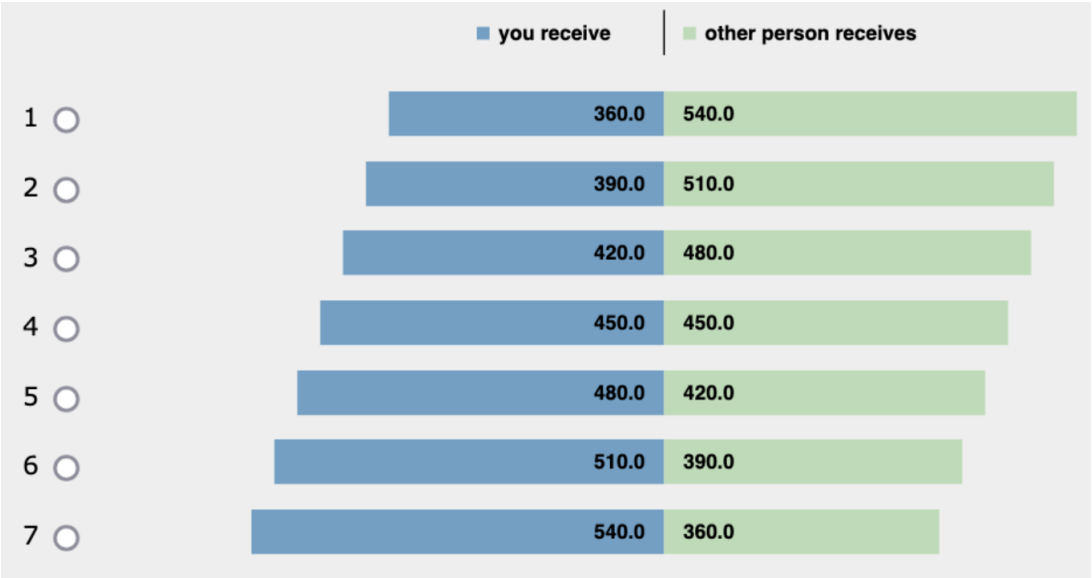
Note. j indicates the unique identifier for each choice situation. Outcomes x and y are expressed in ECUs (experimental currency units), where superscript s refers to *self* and o to *other*. The points (x^s, x^o) and (y^s, y^o) represent the endpoints of the lines shown in Figure 2, with $x^s \geq y^s$. The *domain* categorizes choice situations (see main text). MCR represents the marginal cost of redistribution: positive values indicate that increasing the other's payoff reduces one's own, while negative values indicate that both payoffs increase. α_{crit} and β_{crit} are the critical values for inequality aversion parameters as defined by Fehr and Schmidt (1999).

Respondents chose one option from a set of seven possible bilateral distributions. Each option represented a distribution of monetary payoffs between the respondent *(self)* and an anonymous counterpart *(other)*. The counterpart did not participate in the same allocation decisions, and both parties remained fully anonymous throughout the study. An example choice situation, corresponding to $j = 4$ in Table 1, is depicted in Figure 1, as shown on respondents' screens.

Each choice situation is characterized by two allocation endpoints (x^s, x^o) and (y^s, y^o) where $x^s \geq y^s$, and s and o refer to the payoffs for the *self* and the *other*, respectively. The endpoints define a line in the self-other payoff space, as illustrated in Figure 2.

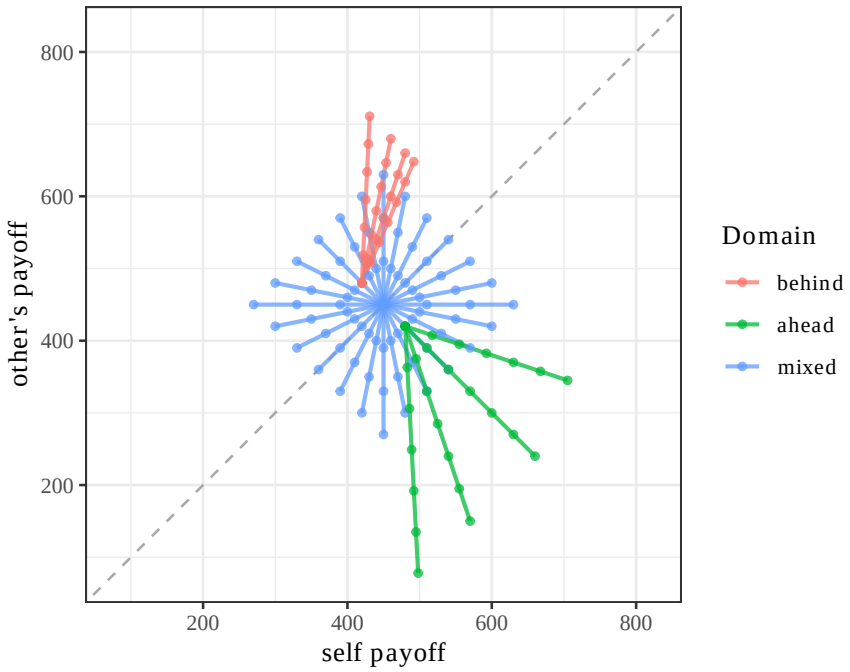
Choice situations vary in the marginal cost of redistribution (MCR), i.e., the money sacrifice *self* has to give up to increase the *others*' payoff by 1 unit—which is inversely related to the slope of the lines shown in the figure—as well as their location within the

Figure 1: Choice interface



Note. In each choice situation (here $j = 4$), respondents were confronted with seven possible distributions between themselves and another person. They were asked to choose one out of the seven options.

Figure 2: Choice situations and domains



Note. Each line corresponds to a choice situation (see Table 1), with the seven dots indicating the choice options displayed on screen. The choice situations (lines) are defined by their endpoints (x^s, x^o) and (y^s, y^o), where $x^s \geq y^s$. The 45-degree line indicates the set of distributions where both individuals are equally as well off (equality line). The lower and upper triangular regions specify the set of distributions where the respondent *(self)* is better and worse off, respectively.

payoff space. We consider three different choice domains: *mixed*, where options allow for higher payoffs to either the self or the other person, *ahead*, where the respondent is always better off than the counterpart, and *behind*, where the respondent is always worse off. In each scenario, respondents were presented with seven equally spaced convex combinations of the two endpoints:

$$\left(z_{ij}x_j^s + (1-z_{ij})y_j^s, z_{ij}x_j^o + (1-z_{ij})y_j^o\right) ,$$

with $z_{ij} \in \{0, 1/6, 1/3, 1/2, 2/3, 5/6, 1\}$ denoting the possible allocation choices, and i denoting the individual index. Higher values of z_{ij} indicate an increase in the respondent's *(self)* payoff.[5] The order of choice situations was randomized for each respondent. Table 1 details the marginal cost of redistribution (MCR), i.e., the amount of the self's payoff that must be sacrificed to increase the other's payoff by one currency unit. It also details the critical values of the Fehr and Schmidt (1999) inequality aversion parameters α and β, where α represents the aversion to disadvantageous inequality (capturing the individual's disutility when he/she receives less than the other), while β represents the aversion to advantageous inequality (capturing the individual's disutility when she/he receives more than the other). Higher values of α and β imply a stronger aversion to being at a disadvantage and advantage, respectively.[6]

2.3 Survey Items

We tested a total of 34 items organized in three sets: *altruism* (11 items), *comparison* (12 items), and *inequality aversion* (11 items). Each of the items within these sets is structured symmetrically. That is, each set includes: (i) a general question about the attitude in focus (labeled as "`Gen`"), (ii) a question addressing the attitude toward strangers (labeled as "`Context`"), and (iii) nine to ten questions describing specific aspects of the attitude (labeled as "`Description`"). The full list of items are in Appendix A.3.[7]

Following Falk et al. (2023), we used their general altruism item as well as their context-dependent item on altruism toward a specific group (in our case, "strangers"). Since their altruism survey module was validated using choice-tasks—where participants selected their preferred donation amount to a charity of their choice—their altruism items related to charity were not applicable to our study, which instead involved the distribution of monetary payoffs between the respondent (*self*) and an anonymous counterpart (*other*). As their guidelines suggest that it is more relevant to tailor the items to the targeted population (p. 1944), we proposed our own list of altruism items that apply in most real-life circumstances—i.e., beyond charitable donations. In particular, our proposed items were inspired from what we believe to be heterogeneous factors that trigger altruistic behavior in daily life, such as disinterested or selfless concern for the well-being

[5] For the situation with $MCR = 0$ (vertical line in Figure 2, corresponding to $j = 7$ in Table 1), $z = 1$ refers, by convention, to the distribution at the very bottom.

[6] The critical values are calculated by equating the slope of the lines with the marginal rate of substitution implied by the indifference curve. Formally, we have:

$$\alpha_{\text{crit}} = -\frac{\text{MCR}}{1+\text{MCR}} \quad \text{and} \quad \beta_{\text{crit}} = \frac{\text{MCR}}{1+\text{MCR}}$$

[7] Our study also included a set of additional survey items that are unrelated to this paper.

of others *("I value the well-being of others more than maximizing my own personal benefit.")*, moral obligation *("I believe that sharing with others, even when not required, is the right thing to do.")*, or personal satisfaction from donating *("I feel fulfilled when I can give something to others, even if it costs me personally.")*.

One limitation we see with altruism items in general is that they seem primarily applicable to situations involving a trade-off between the self's payoff and the other's payoff, as depicted by the negatively-sloped choice situations in the *mixed* and *ahead* domain of Figure 2. To propose items that capture additional considerations for redistribution (including implicit emotions related to such actions), and that are applicable to other types of situations—i.e., including cases where both self's and other's payoff increase or decrease simultaneously, as depicted by the positively-sloped choice situations in the *mixed* and *ahead* domain of Figure 2—we included *comparison* and *inequality aversion* items described as follows.

The comparison items—characterized by the subject's tendency to compare their own situation with that of others—apply to all domains, although they seem particularly relevant for reflecting subjects' behavior in positively-sloped choice situations (Figure 2), i.e., cases where both the self's and the other's payoff increase or decrease in the same direction. For example, while the item *"Whether others have more or less than I do is irrelevant to me"* can refer to all domains, it is specifically designed to capture behavior in the *behind* domain and to assess efficiency (maximizing both payoffs, regardless of the other's payoff). Like the altruism items, we aimed at measuring general attitudes toward comparison *("Do you generally compare what you have with others or not?")* and comparison between the self and the other in anonymous contexts *("Do you generally compare what you have with strangers or not?")*. We then proposed a set of ten descriptive items to relate to realistic scenarios of specific comparison involving some common emotions, such as injustice *("Overall, I feel a sense of injustice when others have more than I do.")*, superiority *("I particularly enjoy situations where I am better off than others.")* and envy *("When I see someone enjoying more resources, I feel a desire to have the same.")*.

Regarding the inequality aversion items—characterized by the subject's sensitivity to unequal distributions between himself/herself and the other—the same logic applied. We aimed to include items that capture a general preference for inequality aversion (e.g., *"Are you generally willing to redistribute resources with others to reduce inequality, or are you not inclined to do so?")*, while tailoring other items to specific cases encountered in certain domains. On the one hand, we proposed items to capture individuals' tendencies toward reducing inequality in advantageous situations, such as *"In situations where I would earn more than others for the same effort, I would feel the need to limit my income at a certain point, even if I could earn more"* and *"I would be willing to sacrifice a large part of my income to slightly reduce that of those less well off than me"*. On the other hand, we proposed items that specifically represent choice situations in the *behind* domain, where individuals are worse off than others: *"In situations where others would earn more than me for the same effort, I would be willing to set an income limit for everyone"* and *"I would be willing to sacrifice a little of my income to drastically reduce that of the most fortunate"*.

2.4 Hypothetical Choices and Real-World Behavior

Building on the Global Preference Survey (GPS) module by Falk et al. (2023), which included hypothetical questions to measure people's attitudes toward risk, time, and altruism (charity), we incorporated a hypothetical version of the incentivized choice experiment. This was done across all domains, i.e., *mixed*, *ahead*, and *behind*, although here we use only the item related to situations involving a trade-off between the self's payoff and the other's payoff. This choice was motivated by its simplicity compared to other scenarios in our set, where the formulation of situations involving simultaneous increases or decreases in both payoffs is more cognitively demanding and complex to articulate.[8] We also used the hypothetical question from Falk et al. (2023) related to charity, which assumes the participant has won $1000 in a lottery and must decide whether he/she would donate a portion of this amount to charity and, if so, in what proportion (Appendix A.4).

Moreover, we included the set of real-world behavior questions of Falk et al. (2023) in the altruism domain, asking about association/volunteering community membership, monthly hours spent volunteering, the number of people the participant knows he/she commits to volunteering, actual donations (whether regular or not), and, if applicable, the amount donated. To also allow for subjective perceptions of inequality in society, which may translate into real-world behavioral support (or activism in its stronger form), we included two Likert-scale questions (0–10) about participants' general concern regarding inequality in the U.S. and their support for inequality-reducing policies (see Appendix A.5 and Epper et al. (2024) for a related item administered to a broad Danish population).

3 Incentivized Preference Measures

In this section we present our findings from the incentivized preference elicitation task. We start with a descriptive analysis of the results (Section 3.1), followed by an exploration of preference types—a qualitative characterization of heterogeneity within our broad population sample (Section 3.2). We then examine the distribution of inequality aversion parameters in our data (Section 3.3).

3.1 Descriptive Results

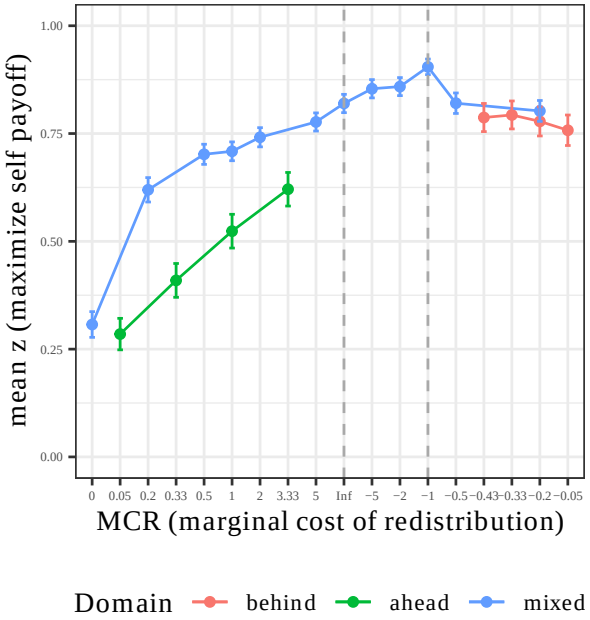
We begin by examining aggregate response patterns within our sample. To do this, we plot the mean z_{ij}-values—where higher z_{ij} indicates a greater propensity to maximize self-payoff—as a function of the marginal cost of redistribution (MCR). Recall that the MCR represents the amount of self-payoff that must be sacrificed to increase the other person's payoff by one currency unit. Again, recall that the MCR is inversely related to the slope of the lines depicted in Figure 2, with $MCR = -1/\text{slope}$.

Figure 3 illustrates the mean responses across the three domains: *mixed* (blue), *behind* (red) and *ahead* (green). In the *mixed* domain, the progression starts with the vertical line in Figure 2 ($MCR = 0$) and moves counterclockwise, transitioning from steeper to flatter negatively sloped lines (corresponding to increasing MCR). As the transition

[8] These additional items with the associated data may be provided upon request.

occurs, redistribution becomes progressively more costly, meaning the self-payoff sacrifice required to increase the other's payoff by one currency unit grows. This trend continues until reaching the choice situation represented by a horizontal line, where the cost becomes infinite (first vertical dashed line in Figure 3). Beyond this point, the lines have a positive slope. Initially, increasing the other's payoff by one unit provides a significant benefit (negative cost) to self, but this benefit diminishes as the slope steepens. At the MCR of -1 (second vertical dashed line in Figure 3), increasing the other's payoff by one unit results in an equivalent benefit for self, maintaining equality. After this threshold, the self-benefit associated with increasing the other's payoff decreases further as the slope continues to steepen. For the *behind* and *ahead* domains, the progression similarly moves from steeper to flatter lines. Figure 3 presents domain-specific responses alongside 95% confidence intervals.

Figure 3: Aggregate response by domain



Note. The figure depicts mean z-values for different marginal cost of redistribution (MCR). The whiskers indicate 95% confidence intervals. The choice situations with $MCR \geq 0$ are represented as negative sloped lines in Figure 2. An infinite MCR indicates a horizontal line in the figure, and a $MCR = -1$ (both persons benefit the same) indicates a line with unit slope.

The results are as follows. In the *mixed* domain, participants are, on average, generous, allocating more to the other person than to themselves. However, as the cost of redistribution increases, participants tend to retain more for themselves, reducing the amount given to the other. Conversely, with increasing benefits of redistribution, they allocate more to themselves, peaking at the point where the total payoff is maximized while maintaining equality. Notably, aggregate behavior does not exhibit perfect maximization of the sum of payoffs at this point. When the other person stands to benefit more than the decision-making participant, individuals demonstrate a willingness to allocate additional resources to the other. In the *ahead* domain—where participants are always better off than the other person—they initially exhibit a willingness to move closer toward equal allocations. However, as the cost of redistribution rises, their willingness to give dimin-

ishes, eventually leading them to move more toward self-payoff maximization. In the *behind* domain, where participants are always worse off than the other person, they generally move toward self-payoff maximization. This behavior aligns with both selfish and efficiency-maximizing motives. As the personal benefit decreases, participants slightly reduce the proportion they retain, but this adjustment is minimal.

Given these aggregate-level results, we now examine behavioral heterogeneity to explain it through variation in survey responses. To achieve this, we adopt two complementary approaches. First, we investigate qualitative differences between participants by identifying preference types based on their response patterns. Second, we explore quantitative differences by estimating individual-level inequality aversion parameters for both the *ahead* and *behind* domains.

3.2 Type Characterization

To identify preference types in our data, we follow Fehr et al. (2024) and search for clusters in the 12-dimensional allocation space. Specifically, each individual is represented as a point in the $z_{.j}$-space, where the allocation in each of the 12 choice situations within the *mixed* domain corresponds to one dimension. We employ the Dirichlet Process (DP) means algorithm (Kulis and Jordan, 2012) with various penalization terms. The penalization term, λ, punishes for the addition of new clusters to the model.[9]

Using this algorithm on the raw data offers several advantages over alternative methods. First, there is no need to commit to a specific behavioral model and a specific error model. Clusters can be identified directly in the allocation space without assuming specific behavioral structures or error models. Second, there is no need to presume the existence of predefined preference types. The algorithm starts with all individuals assigned to a single cluster, represented by the centroid of the mean allocations across all individuals. It iteratively identifies outliers—data points that exceed a predefined threshold (in terms of Euclidean distance)—and creates new clusters as needed. Third, this is a hard clustering algorithm where each individual is assigned to a specific type, producing distinct type labels. This is simpler to interpret compared to probabilistic assignments, as seen in mixture models or related approaches. However, the algorithm does not inherently add interpretation to the resulting clusters. To address this, Fehr et al. (2023) and Fehr et al. (2024) propose three complementary approaches to justify the emergence of three types in their data.

First, the resulting types should exhibit clear qualitative meaning. Fehr et al. (2023, 2024) identify three primary types in Swiss representative samples: one predominantly selfish, one primarily inequality-averse, and one largely altruistic.[10] In this study, we analyze type-specific response signatures to determine whether our results align with these established interpretations. Second, parsimony is a key consideration. A small number of types should explain a large proportion of the heterogeneity in the data. Fehr et al. (2023) find that allowing for a small number of preference types significantly increases precision and out-of-sample predictive ability, while further gains diminish when addi-

[9] The algorithm and the objective function it minimizes are thoroughly described in Fehr et al. (2023).

[10] Similar results emerge when adopting the algorithm to a Danish representative data set. See Fehr and Charness (2025).

tional types are introduced. Their findings suggest that three types represent a "sweet spot" in existing datasets. Third, robustness can be assessed by analyzing how types transition when moving from e.g., two to three, or from three to four clusters. Meaningful types should remain stable within the relevant range of clusters and only lose interpretability when the number of clusters becomes excessively high or low. We confirm this intuition in a robustness exercise detailed in Appendix A.6.[11] These approaches ensure that the preference types identified are both rigorous and interpretable, offering valuable insights into the heterogeneity of preferences in the population.

Building on previous work, we focus on the three-type clustering presented in Table 2. Figure 4 shows that these three types have a clear and unambiguous interpretation. This interpretation aligns with the findings of Fehr et al. (2024) for Switzerland and those reported in Fehr and Charness (2025), based on the data from Epper et al. (2020) for Denmark.

Table 2: Distribution of preference types

Type	Proportion
1	36.25%
2	32.27%
3	31.47%

Note. Proportion of subjects assigned to the three types resulting from employing the DP-means algorithm.

As shown in Table 2, we identify three distinct types, with proportions ranging from 31.5% to 36.3%. The type-specific response patterns, illustrated in Figure 4, provide clear interpretations. Approximately 36% of the sample can be classified as the *predominantly selfish* type, around 32% as the *inequality-averse* type, and roughly 31% as the *altruistic* type.

The predominantly selfish type (Type 1) is characterized by consistently maximizing their own payoff across nearly all choice situations, displaying minimal sensitivity to the cost of redistribution. Notably, a perfectly selfish individual would remain indifferent to all allocations when the cost of redistribution is zero. In our findings, individuals of this type exhibit this behavior even when the cost is only marginally above zero. In the *behind* domain, this type retains as much as possible, with negligible variation in their responses.

The inequality-averse type (Type 2) predominantly selects approximately equal allocations across the board, showing limited sensitivity to the cost of redistribution. For cases where the marginal cost of redistribution (MCR) equals -1, this individual should theoretically be indifferent among allocations, unless they also value the sum of the self's and other's payoff. In our results, this type moves toward more equal allocations in the *ahead* domain, though the tendency is slightly less pronounced compared to other domains.

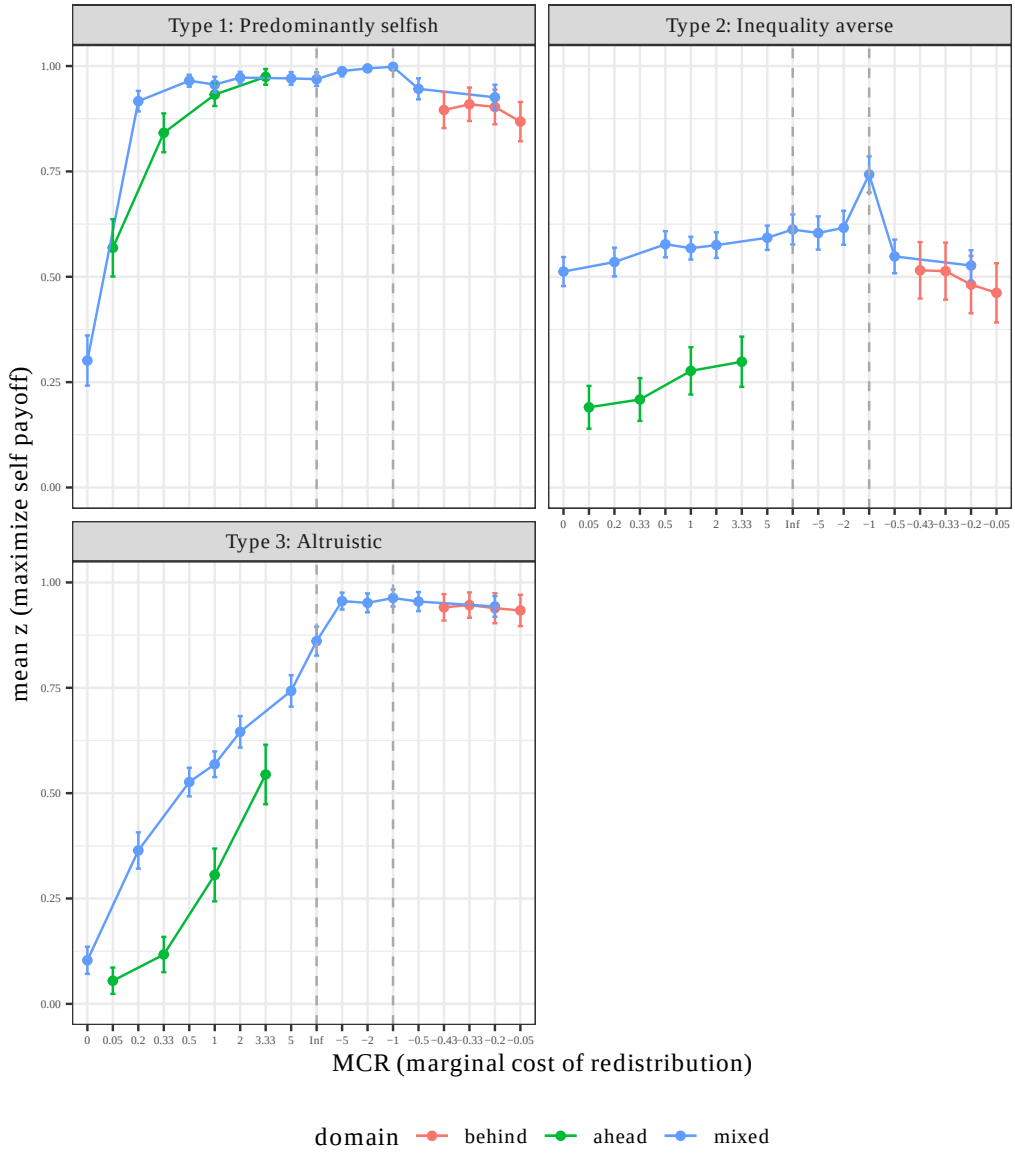
The altruistic type (Type 3) exhibits a strong inclination to allocate substantial resources to the other person, both when they are ahead and when given the opportunity to prioritize the other's payoff over their own. However, in situations where redistribu-

[11] See in particular Figure 12, which depicts type transitions when increasing the number of types.

tion yields mutual benefits (negative MCR, where increasing the other's payoff simultaneously increases their own), this type displays behavior that aligns more closely with selfishness, focusing on maximizing their own gains as well.

Our clustering approach successfully identifies three preference types with interpretations that are consistent with previous findings (Epper et al., 2024; Fehr et al., 2024; Fehr and Charness, 2025). The most notable difference from earlier studies is that, with 36.3%, the selfish type constitutes the largest proportion of our sample, whereas it constitutes a minority in the Swiss samples (between 9.9% and 24%), and a slightly smaller proportion in the Danish sample (32.5%)—see also Fehr et al. (2023).

Figure 4: Type-specific response signatures



Note. The three panels depict mean z-values for different marginal cost of redistribution (MCR) conditional on the preference type. The whiskers indicate 95% confidence intervals.

3.3 Inequality Aversion Parameters

We estimate individual-level parameters of the Fehr and Schmidt (1999) inequality aversion model. When applied to bilateral distributions—the object of choice in our setting—the valuation depends on an individual's own payoff and their relative standing compared to the other person's payoff. The subject's valuation in this setting is expressed as:

$$V((w^s, w^o)) = w^s - \alpha_i \max(0, w^o - w^s) - \beta_i \max(0, w^s - w^o)\,,$$

where w^s denotes the individual's own payoff, defined as $w^s = z_{ij}x_j^s + (1 - z_{ij})y_j^s$, and w^o represents the other person's payoff, defined as $w^o = z_{ij}x_j^o + (1 - z_{ij})y_j^o$. The parameters α_i and β_i are individual-specific preference parameters. The parameter α_i measures inequality aversion when the individual is behind the other person (disadvantageous inequality, or, simply, behindness aversion), while β_i measures inequality aversion when the individual is ahead (advantageous inequality, or, simply, aheadness aversion). The Fehr and Schmidt (1999) model produces piecewise linear indifference curves in the space depicted in Figure 2. The slopes of these curves is closely tied to β_i in the domain where the individual is ahead and α_i where the individual is behind. Table 1 provides the critical values for these preference parameters for each choice situation.

To estimate the model, we assume a random-utility error structure and adopt a method that permits for individual-level heterogeneity. For the error structure, we employ a random-utility framework as introduced by McFadden (1981). Under this model, the probability that individual i chooses alternative k is given by:

$$P_i(k) = \frac{e^{V_{ik}/\lambda_i}}{\sum_m e^{V_{im}/\lambda_i}}\,,$$

where λ_i is an individual-level error parameter representing decision noise. A smaller λ_i implies more deterministic choice behavior. To model heterogeneity, we use a hierarchical Bayesian modeling approach. This approach allows individuals to vary in their preference parameters, α_i and β_i, as well as their error parameter, λ_i. The model constrains individuals with outlier behavior toward the group mean while maintaining flexibility to capture individual differences (partial pooling). Technical details of the estimation procedure are provided in Epper et al. (2024) and Fehr et al. (2023), which estimate such a model to broad population samples from Denmark and Switzerland.

Table 3 presents the sample-level statistics of the estimated parameters. The results indicate that the posterior mean of β exceeds the posterior mean of α, suggesting that aversion to being ahead (advantageous inequality aversion) is stronger than aversion to being behind (disadvantageous inequality aversion). This finding contrasts with the conjecture of Fehr and Schmidt (1999), who proposed $\alpha > \beta$ for their original model. The 95% credibility intervals for both parameters include zero, highlighting substantial variability in inequality aversion across individuals.

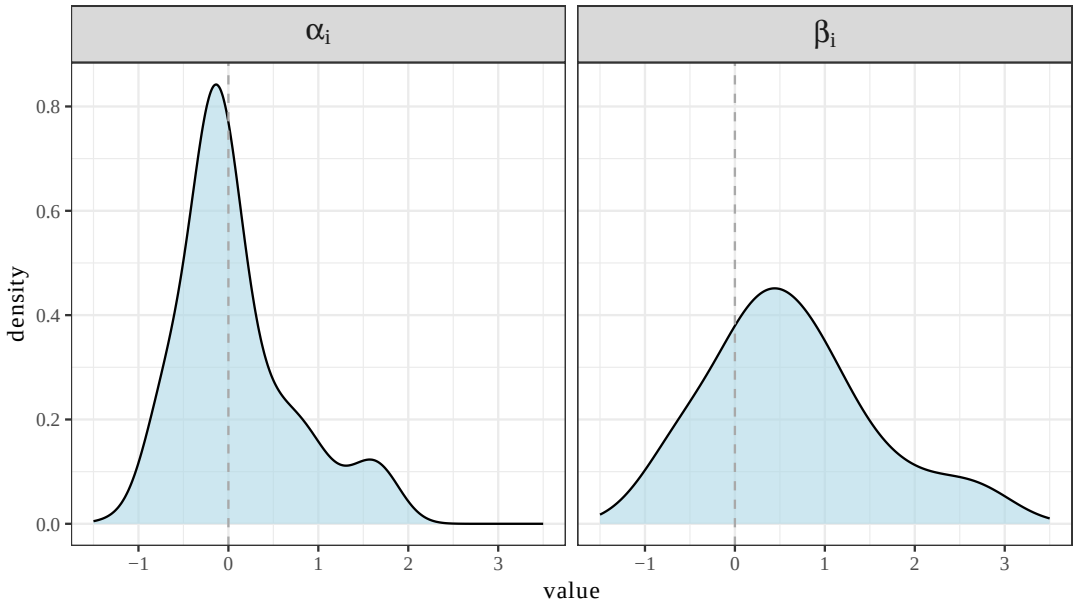
Figure 5 illustrates the distribution of individual-level inequality aversion parameters. The results reveal substantial heterogeneity, with both parameters ranging widely, from slightly negative values to more substantial positive ones.

Table 3: Sample-level statistics

	Estimate	StdDev	2.5%	97.5%
α	0.090	0.634	-0.772	1.698
β	0.636	0.901	-0.851	2.797
λ	0.117	0.122	0.005	0.411

Note. The table lists the means (Estimate), the standard deviation and the 95% credibility interval of the posterior. α and β denote behindness and aheadness aversion, respectively, according to Fehr and Schmidt (1999). λ is the error term in the random-utility specification.

Figure 5: Distribution of individual inequality aversion parameters



Note. The panels illustrate the distribution of behindness aversion (α_i) and aheadness (β_i) in our sample. There is vast heterogeneity in these parameters with an overall tendency toward inequality aversion in both domains.

Table 4: Correlation matrix of model parameters

	α_i	β_i	λ_i
α_i	1.000	0.797	-0.382
β_i	0.797	1.000	-0.308
λ_i	-0.382	-0.308	1.000

Note. The numbers are Pearson correlations between the three individual-level parameters, α_i (behindness aversion), β_i (aheadness aversion) and λ_i (error term). There is a strong individual-level correlation between inequality aversion in the aheadness and the behindness domain.

Table 4 reports the sample level correlations of the parameters across the posterior samples, showing a strong correlation between aheadness and behindness aversion. This relationship is further illustrated in Figure 13 in Appendix A.7, which presents a scatter plot of the two inequality aversion parameters.

In Appendix A.8, we condition the estimates on the types identified in Section 3.2. The results confirm that our interpretation of the response signatures (Figure 4) aligns closely with the structural findings. Appendix A.9 examines the structural model's ability to capture individual-level features of the data. The analysis demonstrates that the model effectively characterizes heterogeneity across individuals and provides accurate predictions of the observed behavioral responses.

4 Survey Responses

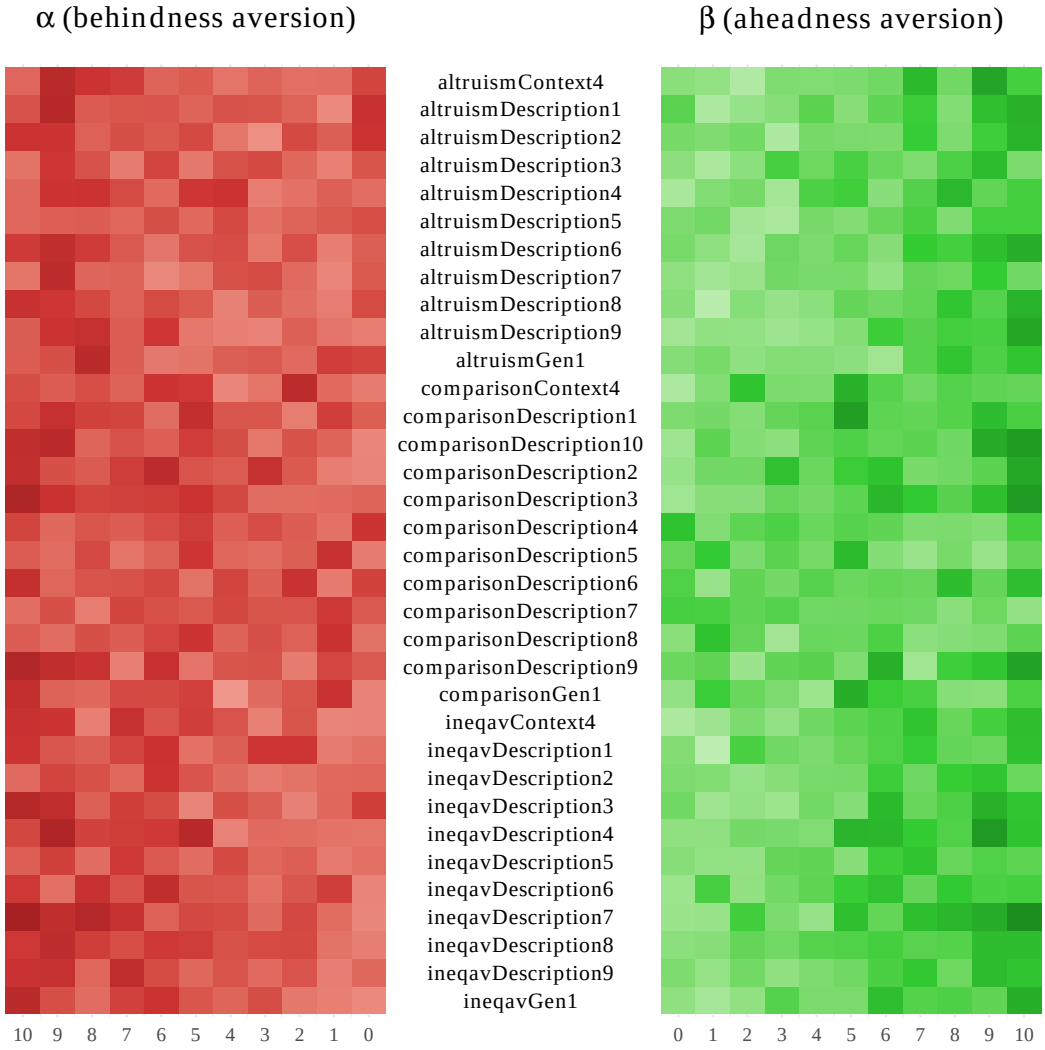
In this section we analyze key survey responses and their association with the behavioral types and their correlation with inequality aversion parameters derived from the incentivized preference elicitation task. We then utilize the 34 survey items as candidate inputs for our predictive model (Section 5). These items span three distinct domains: altruism, social comparison, and inequality. All responses were measured on an 11-point Likert scale ranging from 0 (indicating no adherence to the item) to 10 (indicating extreme adherence to the item)—see Appendix A.3 for detailed formulations of the items and the scale.[12]

The responses to the survey items exhibit substantial heterogeneity, reflecting the diverse perspectives of participants. Figure 6 provides an initial exploration of the relationship between survey responses and the inequality aversion parameters estimated from the incentivized choice task. For each candidate variable, the figure presents a heat map illustrating the association with the two inequality aversion parameters α_i and β_i. Although the associations are not unequivocal for every individual variable, a general pattern emerges: higher levels of inequality aversion (depicted by darker tones in the heat map) tend to correspond to higher response values on the survey items. This suggests a meaningful relationship between self-reported attitudes and the estimated preference parameters we obtained from our incentivized elicitation task.

We also included a question about preferred strategies in a hypothetical scenario

[12]One item, `ineqavDescription3`, was reverse-coded in the original study. For our analyses, its scale was adjusted to ensure that higher values consistently represent less selfish behavior.

Figure 6: Association between 11-point Likert-scale responses and inequality aversion parameters in the 34 survey items



Note. The heat maps illustrate the association between Likert-scale responses and inequality aversion in the *behind* (α) and the *ahead* (β) domain. Darker tones indicate higher degrees of inequality aversion. A smoothing of the parameter values has been applied since some variable feature bins with only a few observations. Overall, there is a tendency of higher degrees of inequality aversion toward higher Likert-scale responses (10). However, there are vast differences across variables.

where participants were asked to decide between the following six options when faced with another participant: (i) take the entire stake (*selfish*), (ii) take more for themselves, but leave some to the other person (*ineqselfish*), (iii) choose an equal allocation (*egalitarian*), (iv) give more to the other person, but keep some to oneself (*inequaltruism*), (v) give the entire stake to the other person (*altruism*), or (vi) select another strategy (*other*) (see Appendix A.4 for the detailed wording).[13] The distribution of responses across these six options was highly uneven, with some strategies (*altruism*, *ineqaltruism*, and *other*) being only rarely chosen (see Table 23 for details). To simplify the analysis, we constructed a binary variable, *anySelfish*, which indicates whether a participant selected a selfish strategy (*anySelfish*=1). Overall, 37.3% of participants opted for a selfish strategy, aligning closely with the proportion of selfish types identified in our clustering exercise.

As shown in Table 5, while these responses contain some predictive signal regarding participants' actual choices, the signal is imperfect, reflecting a notable discrepancy between stated preferences and revealed preferences. Consequently, this survey question appears to offer limited discriminatory power for distinguishing between the two social preference types.

Table 5: Contingency table of subjects stating any selfish strategy vs. the three types identified via clustering

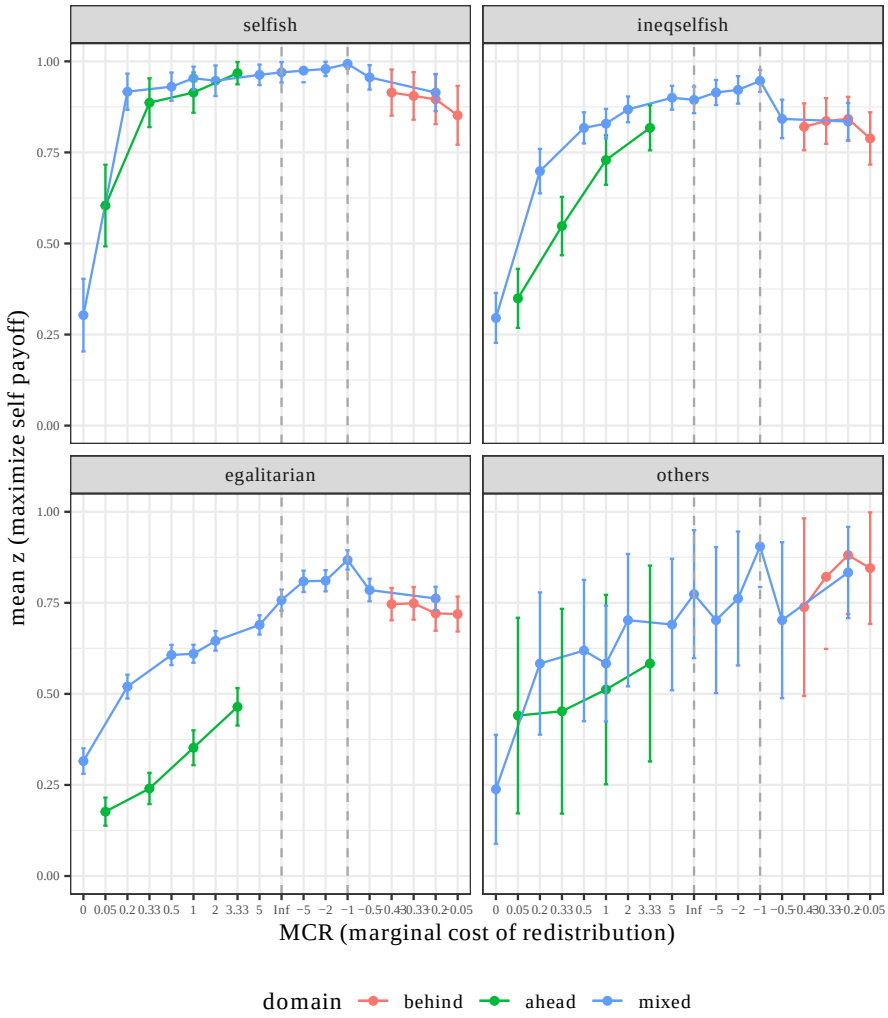
	Type 1: Selfish	Type 2: Inequality averse	Type 3: Altruistic
any selfish strategy	**24.9%**	5.4%	7.0%
other strategy	11.4%	**26.9%**	**24.5%**

Note. The table reports proportions. Stated selfish strategies are indicative for being a selfish preference type as inferred from revealed preference data. However, this signal is far from perfect.

To further assess the effectiveness of strategy responses in predicting allocation choices, consider Figure 7. This figure illustrates the response patterns for four strategy types: (i) participants who chose the fully selfish strategy *(selfish)*, (ii) those who selected a more balanced selfish strategy, taking more for themselves but leaving some for the other participant *(ineqselfish)*, (iii) participants who stated an egalitarian strategy *(egalitarian)*, and (iv) a residual group encompassing other or unspecified strategies *(others)*. Stated strategies are roughly in line with the responses we expect in the different settings of the elicitation task (see also the figure notes).

[13] This survey question appeared at a random point in the survey, either early on, preceding the choice task, or later, following the choice task. We find no evidence that the position of this survey question influenced participants' responses to the task, nor that task responses affected how participants answered the survey question.

Figure 7: Strategy response signatures



Note. Stated strategies are broadly in line with expected (revealed) behaviors. Respondents who stated the *purely selfish* strategy exhibit selfish behavior across the board, the only exception being the area where the cost of redistribution are negligible. Respondents who stated the more balanced strategy of taking more for themselves, but still allocating a smaller part to the other person (*ineqselfish*), reveal a cost-sensitive response pattern. Respondents who stated the *egalitarian* strategy reveal a behavior that is closer to equal allocations, albeit only imperfectly. Finally, respondents who stated one of the *other* strategies reveal a wide variety of behaviors.

5 Survey Module, Scores and Predictions

We develop models to predict type associations and domain-specific inequality aversion parameters. To this end, we tune and train a gradient boosting algorithm on a subset of our data and use it to predict types and inequality aversion for the remaining data.[14] We

[14]More specifically, we utilize the XGBoost (eXtreme Gradient Boosting) algorithm (Chen and Guestrin, 2016)—a regularized gradient boosting algorithm—and perform a grid search on a wide array of hyper-parameters combined with 5-fold cross-validation. This approach minimizes the risk of overfitting while ensuring the model's generalizability and validity.

then evaluate the predictive ability of the large set of survey items using SHapley Additive exPlanations (SHAP values).[15] Our approach identifies a small set of survey items that, when appropriately weighted, explain a reasonably large proportion of behavioral variation we observe in the sample and its subsets.

In what follows, we first develop a classification model to predict assignment to the three preference types identified earlier (Section 5.1). Next, we address the regression problem of predicting aheadness and behindness aversion parameters, estimated from choice data (Section 5.2). Section 5.3 proposes the final survey module items and the scoring method to aggregate the measures, enabling prediction of both types and inequality aversion. We assess the predictive performance of our module using a holdout test set that was not used for any model training and tuning. In Section 5.4, we demonstrate that our proposition is able to explain variation in stated hypothetical and real-world settings where inequality aversion and altruism are expected to play a role.

5.1 Predicting Types

Our first primary objective is to develop a model capable of predicting whether an individual exhibits a preference type classified as *selfish*, *inequality averse*, or *altruistic*. This prediction relies on a comprehensive set of survey items designed to capture relevant dimensions of social attitudes. Specifically, we utilize survey items assessing *altruism*, *social comparison*, and *inequality aversion*, all of which were included in our survey. We also incorporate responses to the *strategy* question, where participants articulated their qualitative preferences in a hypothetical decision-making scenario.

We proceed as follows. We split our data into a training set of 402 respondents (roughly 80% of the full sample) and a holdout test set of 100 observations (roughly 20%). As its name suggest, the training set is used for the training and tuning of the model. We make use of the holdout test set later, where we use it to assess the model's performance on data it has not seen before. To optimize the predictive performance of our model and make efficient use of our (training) data, we further employ a 5-fold cross-validation. In this procedure, the dataset (the 80% of the full sample) is partitioned into five approximately equal subsets, or "folds". The model is then iteratively trained on four folds (the training set of a fold) and tested on the remaining fold (the validation set of a fold). This process is repeated five times, with each fold serving as the validation set exactly once. Based on these five iterations, we compute performance metric as an average of the individual metrics. In the classification problem we study in this part, our performance metric is the *accuracy*, i.e., the proportion of correctly predicted type labels, and our objective is a softmax function.

We use this cross-validation procedure to tune the gradient boosting model's hyperparameters, i.e., parameters that specify the way the model learns from the data, using a

[15] We deliberately exclude other variables, such as socioeconomic or political background, as potential predictors. Our primary objective is to develop a module that serves as a substitute for incentivized preference measures with relatively high accuracy, rather than to construct a predictive model that leverages all available data to forecast preference types or inequality aversion. In Section 5.4, we demonstrate a typical application of our approach: in regression analyses, we replace the preference measures with an index measure, *while* still controlling for socioeconomic and other explanatory variables.

grid search over a wide set of tuning parameters. Specifically, these parameters contain the number of trees, their depth, the learning rate, the minimum loss reduction, the fraction of features (variables), the minimum sum of weights, and the fraction of data used for the boosting. For each hyperparameter combination, we perform the 5-fold cross validation and compute the average accuracy. We then chose the best set of parameters for our final model. This selection procedure minimizes the risk of overfitting and ensures a balanced evaluation of model performance. Importantly, it enhances the model's generalizability by enabling robust predictions on data it has not encountered during the training steps.

To assess our final model's performance, we make use of our holdout test set. Table 6 shows the confusion matrix for these out-of-training-set predictions. It compares the predicted type assignment with the actual (reference) type assignment we obtained via the clustering exercise. Note that we have exactly 100 respondents in this sample. Hence, the numbers can be directly interpreted as percentages.

Table 6: Confusion matrix for holdout test set | Full model

Actual / Predicted	Type 1	Type 2	Type 3
Type 1	27	4	5
Type 2	9	14	8
Type 3	7	10	16

Note. The contingency table (or confusion matrix) reports on how many respondents were correctly or incorrectly assigned to one of the types in the holdout test set. Note that we have exactly 100 respondents in the holdout test set, such that the numbers can be interpreted as proportions of correct/incorrect predictions per bin.

With 57% of correctly predicted classifications, the accuracy of our model is relatively high. This accuracy represents a substantial and significant improvement over the no-information rate (NIR), which simply uses the largest preference type in the holdout set as the prediction (43%). A statistical test confirms the significance of this improvement: The p-value for comparisons of the accuracy with the NIR lies below 0.01, indicating strong evidence of the model's predictive capability. To evaluate misclassification patterns, we employ McNemar's test, which assesses whether significant differences exist between false assignments. The high p-value of 0.48 suggest that misclassification patterns are stable, further supporting the model's reliability.

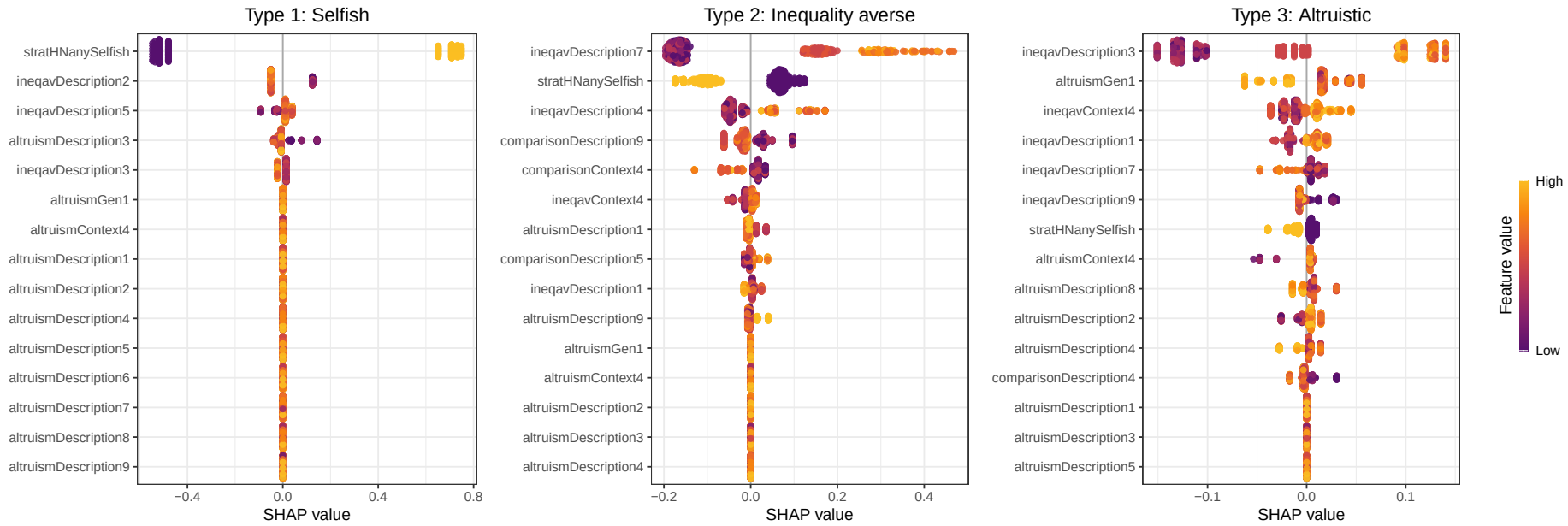
The model performs particularly well in distinguishing *selfish* individuals from *non-selfish* individuals, exhibiting high sensitivity and detection rates for the *selfish* type (Type 1). This result aligns with expectations, as *selfishness* tends to correspond to more distinct and measurable patterns in survey responses. In contrast, differentiating *inequality averse* individuals from *altruistic* individuals presents greater challenges. This difficulty likely stems from the nuanced and overlapping characteristics of these preference types, which may be driven by similar underlying motivations and reflected in comparable survey response patterns. To further investigate these challenges and gain a deeper understanding of the importance and discriminatory ability of variables, we analyze SHAP values derived from the calibrated model. They quantify the contribution of each variable (feature) to the model's type predictions, providing both global and local interpretabil-

ity. These values are particularly useful for identifying key predictors and understanding how individual survey items impact the classification of preference types—which is the main objective in this exercise.

Figure 8 displays SHAP values for each of the three preference types as computed from the full dataset. For each type, the most predictive variables (features) are listed from top to bottom in order of their overall importance (computed as the mean absolute SHAP value). Positive SHAP values indicate a contribution toward predicting assignment to that type, while negative values indicate a contribution away from predicting that preference type. Each point represents an individual data point for a specific variable. The color of the points (heat) correspond to the variable value (yellow for high values, purple for low values).[16] Looking at the points, we can thus see how variables affect SHAP contribution. A wider spread of the data points for a given variable indicates that the variable's impact on the prediction varies significantly across observations. For reference, Figure 9 also presents the mean absolute SHAP value for the three types. This figure gives a quick indication on the importance of different variables (features) in predicting assignment to the different types. If the ordering of the values is unbalanced across types, this suggests that a variable has discriminatory power to separate between preference types. This is of particular relevance to disentangle the two (harder to distinguish) non-selfish types (Type 2 and 3).

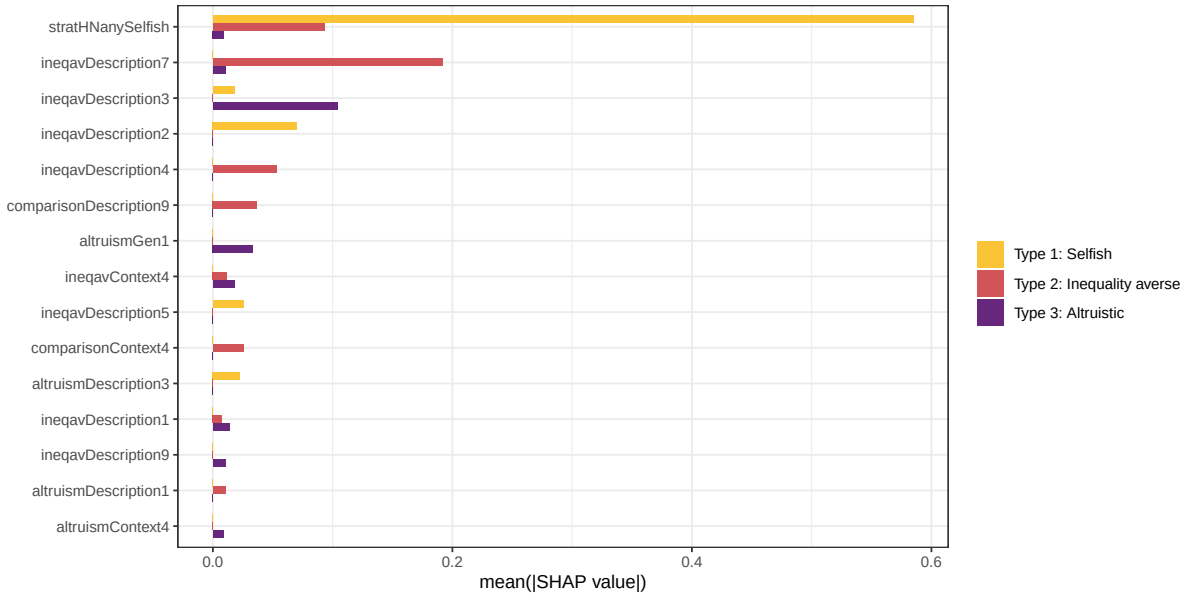
[16] Recall that our strategy variable `stratHNanySelfish` is a binary variable with a value of 1 indicating a *selfish* strategy and a value of 0 indicating a *non-selfish* strategy.

Figure 8: SHAP values by type



Note. The beeswarm plots show the SHAP values for the variables (features) of highest importance separately by preference type. The hypothetical strategy question (`stratHNanySelfish`) discriminates well between selfish (Type 1) and inequality aversion (Type 2). However, it is less powerful in identifying altruistic types (Type 3). The survey item `ineqavDescription3` performs particularly well in identifying altruism (Type 3), followed by `altruismGen1`. Variables that have little to no predictive power are omitted.

Figure 9: Mean absolute SHAP values by type



Note. The figure lists the top 15 predictors ranked by their importance computed from mean absolute SHAP values. The variables (features) work differently well in predicting type assignment. For instance, `stratHNanySelfish` is highly predictive for *Type 1: Selfish*, and to some extent for *Type 2: Inequality averse*. However, it has little to contribute for identifying *Type 3: Altruistic*. `ineqavDescription7` and `ineqavDescription3`, on the other hand, perform comparatively well in predicting *Type 2: Inequality averse* and *Type 3: Altruistic*, respectively.

The figure provides key insights into how individual variables contribute to type assignment in our model. Notably, the strategy variable `stratHNanySelfish` consistently stands out as one of the top predictors. This variable plays a crucial role in distinguishing between the selfish and the inequality averse type. In particular, endorsing a selfish strategy significantly increases the probability of being classified as the selfish type (Type 1), while simultaneously reducing the likelihood of being categorized as the inequality-averse type (Type 2). Although its predictive power for the altruistic type (Type 3) is less pronounced compared to the other two types, it still exhibits clear, albeit relatively weak, discriminatory power (it is the 7th most important predictor for this type only). These findings are consistent with intuitive expectations. Analyzing the mean response patterns of the altruistic type (see Figure 4) reveals that altruistic individuals demonstrate cost-sensitive giving when they are ahead of others, while their behavior aligns more closely with selfishness when they are behind others.

Survey variables capturing respondents' willingness to distribute money between themselves and another person across various scenarios (the `ineqavDescription` items) also rank prominently among the top predictors. These variables, while to some extent related to the strategy question, provide more detailed insights due to their framing and use of an 11-point Likert scale (see, for example, `inequavDescription3`, `inequavDescription7`, and `inequavDescription2`). Interestingly, responses to the general altruism item, `altruismGen1`, stands out for its contribution to distinguishing altruists (Type 3) from non-altruists (Type 1 and 2). This suggests that certain survey items capture essential behavioral nuances specific to altruistic tendencies, reinforcing the value of these variables in enhancing the model's classification accuracy.

Despite the strong predictive performance of the aforementioned variables, a considerable subset of the survey items contributes minimally to the model's predictive ability. These low-impact variables can be identified and excluded from the model without compromising much of its predictive power. We discuss this in more detail in Section 5.3.

5.2 Predicting the Degree of Inequality Aversion

In this section, we extend our analysis by studying the prediction of inequality aversion parameters α_i (capturing behindness aversion) and β_i (capturing aheadness aversion) as conceptualized in Fehr and Schmidt (1999). These parameters were estimated for each respondent in our dataset. Our objective is to explore the extent to which variation in these parameters can be explained by the comprehensive set of survey items included in our study. Specifically, we aim to identify survey items that are most predictive of higher values in the sample distributions of α_i and β_i.

This regression-based analysis represents a more complex task compared to the classification problem explored earlier. Here, the goal is not to precisely predict the numerical values of the inequality aversion parameters but rather to determine whether variation in these parameter values can be systematically predicted using the available survey data. Additionally, we aim to evaluate whether predictability differs across domains: different variables may have bite in predicting people's dislike of being behind (α_i) or being ahead (β_i). As previously observed, the distribution of α_i is more concentrated compared to that of β_i, indicating less heterogeneity in behindness aversion (see Figure 5). This reduced variability likely makes it more challenging to explain differences in α_i, whereas the greater heterogeneity observed in β_i should facilitate better predictability in the domain of aheadness aversion. Using the same methodological framework as in the classification analysis—a 5-fold cross-validation jointly with a grid search for the hyperparameter tuning—we train a separate model for each inequality aversion parameter. As these are regression models, our objective function has to be adopted to a squared error. Moreover, we use the root mean square error as our metric for choosing the best set of hyperparameters. To evaluate our models, we again assess their performance in predicting outside of the training sample, i.e., in the holdout test set.

In line with our expectation, the performance of the model differs substantially between the two inequality aversion domains. The model aimed at predicting β_i (aheadness aversion) performs substantially better than the model aimed at predicting α_i (behindness aversion): The proportion of the variation in β_i that is predicable from the variables (22.67%) is considerably higher than the proportion of the variation in α_i (7.77%). However, we see a significant association between the ranks of predicted inequality aversion parameters and those we observe in our data. This is the case for both models (domains), albeit the association is arguably stronger for β_i than for α_i.

To interpret the contributions of individual variables to the models, we again examine the SHAP values computed from the full dataset. Figure 10 reveals that similar variables are among the top predictors for both α_i and β_i. Notably, respondents who indicated a selfish strategy in the hypothetical scenario are systematically predicted to have lower values for both α_i and β_i, suggesting that selfish strategies are associated

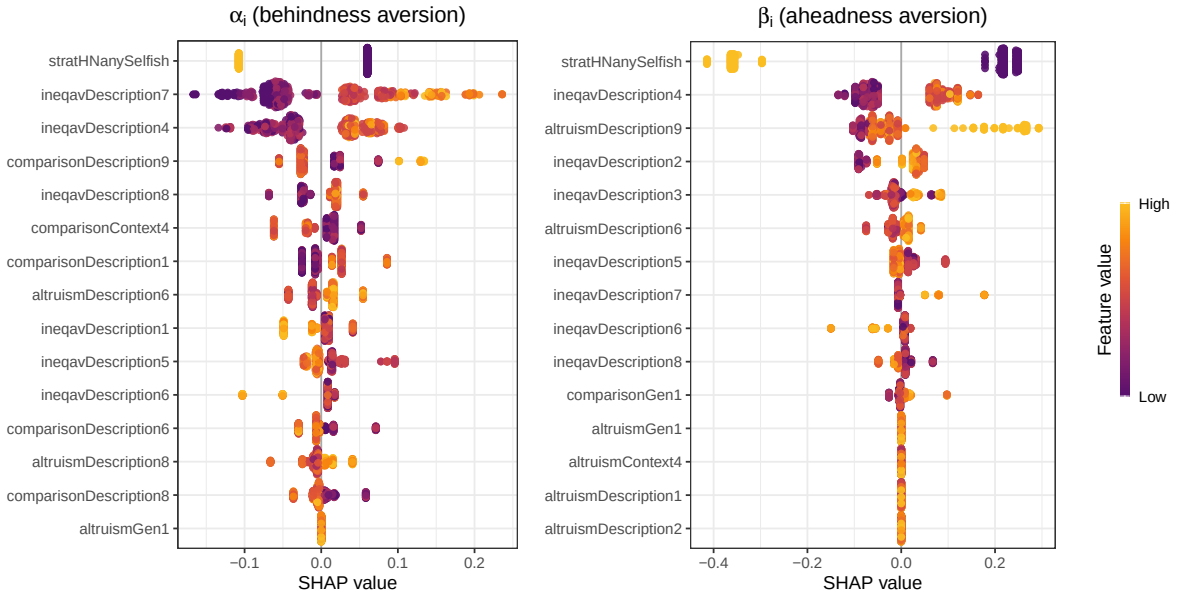
Table 7: Model performance in the holdout test set using different metrics | Full models

Model	RMSE	R^2	MAE	ρ (p-value)
α_i (behindness aversion)	0.646	7.77%	0.501	0.242 (0.015)
β_i (aheadness aversion)	0.866	22.67%	0.661	0.511 ($\approx$ 0)

Note. The table reports several metrics on the performance of the regression models when predicting the inequality aversion parameters in the holdout test set. RMSE is the root mean square error. The coefficient of determination, R^2, has the usual interpretation. It states how much of the total variance is explained by the model. MAE is the mean absolute error. The last column lists the Spearman rank correlation coefficients (ρ) and the p-values of the corresponding test on the association between predicted vs. observed parameter values. What is striking is the better performance of the model aimed at predicting aheadness aversion (β_i) as opposed to the model aimed at predicting behindness aversion (α_i).

with reduced concern to inequality in both domains. Among the survey variables, the inequality description items (particularly `ineqavDescription4`, `ineqavDescription7`, and `ineqavDescription9`) again emerge as important predictors on our list. These variables provide valuable insights into respondents' attitudes toward inequality and their sensitivity to distributional preferences, making them central to the predictive models for both α_i and β_i.

Figure 10: SHAP values by inequality aversion parameter

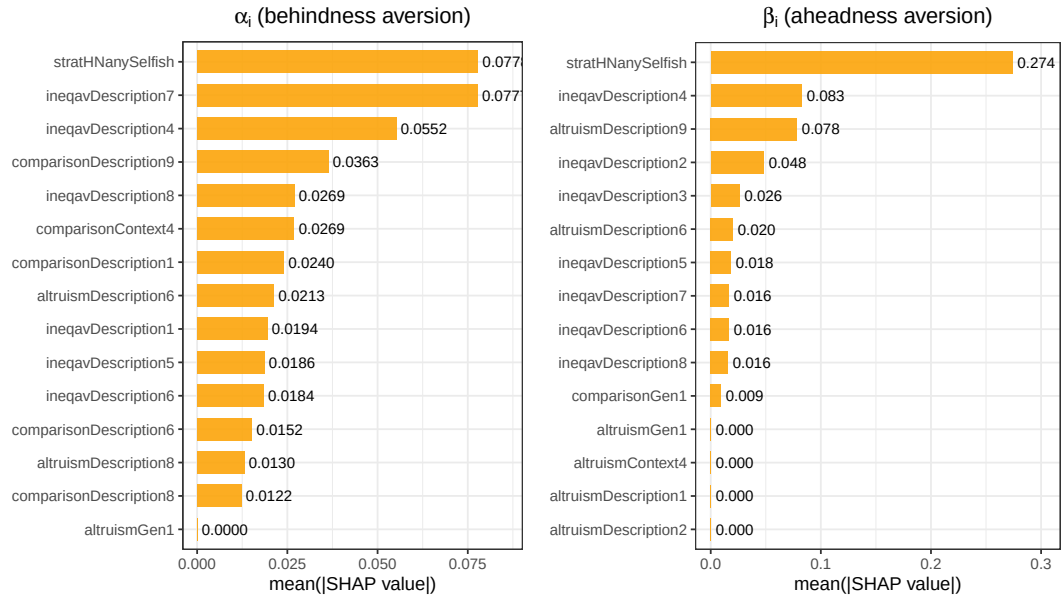


Note. The beeswarm plots show the SHAP values by inequality aversion parameter based on two separate models. Stating a selfish strategy is associated with lower α_i and β_i values. The survey item `ineqavDescription4`—which is a similar type of question as `ineqavDescription3` (see the type SHAP plots)—is the next best predictor for both aheadness and behindness aversion. Only the top 15 predictors are displayed.

For completeness, we again provide the figures listing the mean absolute SHAP values for the top predictors in the models. Figure 11 shows that the strategy item, `stratHNanySelfish`, stands out in the aheadness aversion model. The remaining predictors have mean absolute SHAP values that are in the ballpark of those in the behindness

aversion model, however. Only 14 (11) variables feature positive mean contributions for α_i (β_i).

Figure 11: Mean absolute SHAP values by inequality aversion parameter



Note. The two panels lists the top predictors ranked by their importance computed from mean absolute SHAP values separately for the two parameter regression models. `stratHNanySelfish` is the best predictor in both models. However, in terms of quantitative contribution it does much better in the β_i (aheadness aversion) model. This variable is followed by some key variables we already identified in the classification exercise (in particular `ineqavDescription4` and `ineqavDescription7`).

5.3 Constitution of The "Hearts-and-Minds" Module

Using the importance rankings of the predictors identified in the previous sections, we can now select a concise set of survey items that, when appropriately weighted, provide reasonably good quality predictions for type assignment and individual heterogeneity in inequality aversion. To maintain brevity and focus, we select the top two predictors (features with the highest SHAP values) from each of the three models with outcome variables *type assignment*, *aheadness aversion*, and *behindness aversion*. For the classification model, we include the two top predictors across all three preference types. This approach results in a total of seven survey items, which, we argue, strike an effective balance between brevity and predictive validity. These seven items take approximately 1 to 2 minutes to administer, making the module practical for integration into larger surveys or field studies.

Table 8 presents the full set of seven survey items that constitutes the *"Hearts-and-Minds"* module. All items have high feature importance (in terms of SHAP values) in our original models and/or they provide good discriminatory power between preference types or parameters.

Table 8: The "Hearts-and-Minds" survey module

Item	Description
strategyHN	*"Imagine you are in a situation where you have to distribute money between yourself and an anonymous person. Neither of you will see or interact with the other. You have absolutely no information about the other person's circumstances (such as his/her wealth). The only thing you know is that nobody, except you and the other person, will ever know your choice. What would you do? I would..."*
ineqavDescription7	*"I would be willing to sacrifice a large part of my income to slightly reduce that of those less well off than me."*
ineqavDescription3	*"If I have the choice to distribute resources with strangers, I would rather keep more for myself and give less to others."*
ineqavDescription2	*"I would prioritize equity over maximizing my own benefits if I were in a situation where I had to distribute resources with others."*
ineqavDescription4	*"If I have the choice to distribute resources with strangers, I would rather give more to others and keep less for myself."*
ineqavDescription9	*"I would be willing to sacrifice a little of my income to drastically reduce that of the most fortunate."*
altruismGen1	*"Are you generally willing to share with others without expecting something in return, or are you not willing to do so?"*

Note. For the strategyHN item the answer categories are: (i) *"keep everything for myself"*, (ii) *"take a larger portion for myself and leave a smaller portion for the other"*, (iii) *"make an approximately equal distribution between myself and the other person"*, (iv) *"take a smaller portion for myself and leave a larger portion to the other person"*, (v) *"give everything to the other person"*, (vi) *"do something else (see below)"*. A single option must be selected. The last option of the strategy item is followed by an open text field. The stratHNanySelfish variable we use is a dummy variable which is 1 if option (i) or (ii) was selected and 0 otherwise. The 11-point Likert scales are as follows. For the *ineqavDescription* items: *"0: does not describe me at all"* to *"10: describes me perfectly."* Note again that we intentionally reverse-coded *ineqavDescription3* to check participants' consistency in responses. For the altruismGen1 item: *"0: completely unwilling to do so"* to *"10: very willing to do so."*

The strategyHN item acts as a clear indicator of selfish behavior, functioning as a dummy variable that strongly associates with the selfish type (Type 1) and decreases the likelihood of being classified as Type 2 (inequality averse) or Type 3 (altruistic). Similarly, this variable effectively predicts low values of both behindness and aheadness aversion parameters. The ineqavDescription3 and ineqavDescription4 items are identical in terms of consequences but are framed differently. While ineqavDescription3 adopts a rather self-interested approach, focusing on personal endowment ("*keep more to myself*"), ineqavDescription4 emphasizes the other ("*give more to others*"). This dual framing—focusing on "keeping" *versus* "giving"—ensures that the module does not overlook individuals who may express different preferences depending on how the situation is framed. This duality addresses a known issue in decision-making under framing effects (Tversky and Kahneman, 1981), allowing participants to more clearly articulate a "reflective" preference when they are aware of both frames (Lecouteux and Mitrouchev, 2024). Note that ineqavDescription7 and ineqavDescription9 are also joined. There is no dual framing here (due to different consequences), but they encapsulate the willingness to sacrifice to reduce inequality in terms of the amount (large/small) and the

targeted population (poor/rich). For real-world cases where individuals have political opinions about what to do with the most poor and/or fortunate, these items can be particularly useful. `ineqavDescription2` prompts respondents to consider how they balance their own personal benefits with a concern for fairness and equality when allocating resources. This item captures the underlying tension that individuals might feel between self-interest and the desire to ensure equal opportunities or outcomes for others. By addressing this cognitive *trade-off*, the item helps to gauge whether individuals prioritize social equity over their personal advantage in decision-making scenarios. As for `altruismGen1`, it is Falk et al. (2018, 2023)'s simple item, which performs well in their GPS module. They find a coefficient weight resulting from their OLS regressions of 0.635 (Falk et al., 2018) and 0.3210 (Falk et al., 2023), although with a different social preference elicitation than ours (charity donation in their case). Note also that we modified the phrasing of Falk et al. to better fit with one's general (i.e., a-contextual) tendency to give. The original phrasing of Falk et al. is: "*How willing are you to give to good causes without expecting anything in return?*" (we emphasize "good causes"). This item particularly captures altruism in a disinterested/Kantian sense ("without expecting something in return"), aligning with the most common understanding of what *pure* altruism entails.

Note that none of the `comparison` items are retained in our final survey module due to their poor predictive ability. This contrasts with our prior intuitions, as α and β parameters are inherently dependent on a *comparison* between one's payoff and that of the other person. This relationship is notably emphasized in Fehr and Schmidt (1999), who emphasize that "fairness judgments are inevitably based on a kind of neutral reference outcome" (p. 820).[17]

In summary, our module is constituted by one hypothetical item (`strategyHN`), one trade-off item (`ineqavDescription2`), one simple and powerful item (`altruismGen1`) that has already proven effective in prior studies (Falk et al., 2018, 2023), two "dual framing" items (`ineqavDescription3` and `ineqavDescription4`), and two additional items that explore attitudes toward the rich and the poor with different stakes (`ineqavDescription7` and `ineqavDescription9`).

Based on our proposed module, we now train and fine-tune three *reduced* models: one with the aim to predicting preference types, and two with the aim to predicting aheadness and behindness aversion, respectively. We again train these models on our testing dataset to ensure robust performance, and, then, document these reduced models' ability to predict types and parameters within the holdout test set. Table 9 illustrates that the reduced model is even doing slightly better than the full model (see the confusion matrix in Table 6).

The accuracy of the reduce model is with 63% about 6 percentage points higher than in the full model. This is substantially and significantly higher than the no-information rate (NIR) of 43%. This increased accuracy, however, comes at the cost of a bit more systematic misclassification. With 0.25, the p-value of a McNemar's test, however, is still beyond any level of significance. Moreover, the model has discriminatory power between Type 2 (inequality averse) and Type 3 (altruistic), albeit its ability to distinguish these

[17]The reference point in the social domain could be elicited in further research, if not already undertaken—see Baillon et al. (2019) in the risk domain.

Table 9: Confusion matrix for holdout test set | Reduced model

Actual / Predicted	Type 1	Type 2	Type 3
Type 1	32	4	4
Type 2	5	19	13
Type 3	6	5	12

Note. The contingency table (confusion matrix) reports on how many respondents were correctly or incorrectly assigned to one of the types. Note that we have exactly 100 respondents in the holdout test set, such that the numbers can be interpreted as proportions of correct/incorrect predictions per bin.

types is—without surprise—far from perfect. We conclude that—even with our relatively compact survey module—we are able to predict preference types with an accuracy far beyond chance.

Examining the inequality aversion parameters, the reduced model benefits from its smaller set of predictors, potentially mitigating some overfitting challenge we faced in the full model. It demonstrates superior generalizability and portability in predicting considerably better in the holdout test set. Table 10 reports the different metrics. The coefficients of determination (R^2) are better for both reduced models as compared to their full model counterparts. Most strikingly, however, the behindness aversion model features a substantially better ability in predicting out of sample, indicated by an R^2 close to the aheadness aversion model and rank correlations that lie in a similar region (coefficients of 0.457 and 0.545, respectively, and p-values of ≈ 0).

Table 10: Model performance in the holdout test set using different metrics | Reduced models

Model	RMSE	R^2	MAE	ρ (p-value)
α_i (behindness aversion)	0.580	27.59%	0.440	0.457 (≈ 0)
β_i (aheadness aversion)	0.833	29.54%	0.640	0.545 (≈ 0)

Note. The table reports several metrics on the performance of the regression models when predicting the inequality aversion parameters in the holdout test set. RMSE is the root mean square error. The coefficient of determination, R^2, has the usual interpretation. It states how much of the total variance is explained by the model. MAE is the mean absolute error. The last column lists the Spearman rank correlation coefficients (ρ) and the p-values of the corresponding test on the association between predicted vs. observed parameter values. For the reduced models, the model aimed at predicting aheadness aversion (β_i) only performs similarly well as than the model aimed at predicting behindness aversion (α_i).

Our reduced models thus perform well in predicting both preference types and inequality aversion differences, even though they are based on a relatively concise survey module with seven items. Practitioners who incorporated our module in their survey can use the boosting weights of our three models to obtain predictions. Instructions on how to load the model and predict types and individual heterogeneity is available here: `https://gitlab.com/thomasepper/repl-MEL-surveyModule`.[18]

[18]Note that the predictive performance of our models may be further improved by retraining the model on more diverse data sets that include our survey module and real-incentivized preference elicitation tasks.

5.4 External validity

We eventually examine the predictive power of our scores in relation to both actual (stated) and hypothetical behaviors that are expected to be associated with inequality aversion and altruism. To this end, we present the results of a selected set of regressions. Additional details, including the bivariate associations between these variables and our preference measures, as well as results from a series of supplementary regressions, are provided in Appendix A.10.

To assess the explanatory and predictive capabilities of our module scores for behavior, we first compute these scores across the full dataset. Subsequently, we regress the stated behavioral variables on the scores and, for comparison, on the estimated aheadness and behindness aversion parameters from the incentivized preference elicitation task. Our analysis focuses on four key behavioral variables: (i) support for redistributive policies, (ii) engagement in volunteering, (iii) hours spent volunteering, and (iv) willingness to donate to charity following a windfall.

Motivated by Fehr et al. (2024), we begin by examining support for redistributive policies. Similar regression analyses have been conducted by Epper et al. (2024), who investigated the relationship between inequality aversion parameters—estimated from an incentivized preference elicitation task—and support for public policies. Their study also incorporates a comparable set of control variables to assess the robustness of their findings. However, it is important to note that differences in their outcome scale and control variable specifications limit direct comparability of effect magnitudes between their results and ours.

Table 11 presents the key regression results for our policy support variable, `ineqPolicy`. It is measured on an 11-point Likert scale, where higher values indicate greater support for redistribution (for the exact wording of the question, see Appendix A.5). The table reports results for four models. Model (1) and Model (1c) use the individual behindness aversion (α_i) and aheadness aversion (β_i) parameters estimated from the incentivized preference elicitation task, without and with the inclusion of a comprehensive set of control variables, respectively. Model (2) and Model (2c) follow the same structure but replace the estimated parameters with scores derived from our survey module. In all models, we use percentile ranks of the preference parameters as regressors. The control variables include indicators for income class, education level, age, gender, immigration status, marital status, and the presence of children living in the household. Specifically, marital status is captured using dummies for being married, divorced, separated, or widowed. The intercept represents the baseline support for redistribution for an 18-year-old, non-immigrant male with median income, a high school degree, no marital history (neither married, divorced, separated, nor widowed), and no children living in the household.

The regression results reveal a significant positive association between aheadness aversion (as reflected by β_i and the related survey module-based score) and support for redistributive policies. The coefficients remain relatively robust when the full set of control variables is included. Notably, the relationship between preferences and policy support is stronger when using the survey module based scores compared to the estimated parameters. Epper et al. (2024) report similar findings regarding aheadness. However,

Table 11: Regression Results for Support for Redistributive Policies (`ineqPolicy`)

Variable	(1) estimated	(1c) estimated	(2) score	(2c) score
behindness av.	-0.293 (0.667)	-0.06 (0.672)	0.01 (0.72)	-0.216 (0.726)
aheadness av.	1.52 (0.667)**	1.214 (0.669)*	2.374 (0.73)***	2.357 (0.732)***
Intercept	6.237 (0.292)***	7.823 (0.731)***	5.704 (0.28)***	7.286 (0.733)***
Controls	no	yes	no	yes
R^2	0.015	0.117	0.048	0.144

Note. The response variable was measured on an 11-point Likert scale ranging from 0 to 10, with 10 indicating the highest support. Model (1) and (1c) use the inequality aversion parameters *estimated* from the incentivized preference elicitation task without and with controls, respectively. Model (2) and (2c) substitute the inequality aversion parameters with our survey module-based *scores*. The controls include age, immigrant status, income class, highest degree of education, civil status and a dummy for children in the household. The intercept (baseline level) refers to an 18 year old, non-immigrant male with median income and a high school degree who is neither married, nor divorces, separated or widowed, and has no dependent children. p-values: $0 \leq^{***} < 0.01 \leq^{**} < 0.05 \leq^{*} < 0.1$.

their analysis also identifies a significant association between behindness aversion and policy support, a relationship we do not observe in our data. Given the use of similar elicitation methods and the same estimation protocol in both studies, this discrepancy is likely attributable to difference in sample characteristics (U.S. representative vs. Danish representative sample).[19]

To further assess the external validity of our survey module-based scores, we analyze a set of survey questions proposed by Falk et al. (2023), in which respondents report their volunteering activities. Table 12 presents the results from a linear probability model where a binary variable indicating volunteering is regressed on the preference parameters, both without and with the inclusion of control variables.

Table 12: Regression Results for Volunteering (`socialMember1`)

Variable	(1) estimated	(1c) estimated	(2) score	(2c) score
behindness av.	0.257 (0.096)***	0.209 (0.095)**	0.323 (0.104)***	0.257 (0.103)**
aheadness av.	-0.058 (0.096)	-0.015 (0.095)	0.052 (0.105)	0.087 (0.103)
Intercept	0.175 (0.042)***	0.22 (0.103)**	0.089 (0.04)**	0.127 (0.104)
Controls	no	yes	no	yes
R^2	0.02	0.159	0.055	0.184

Note. The response variable is binary. The reported results are for a linear probability model. Model (1) and (1c) use the inequality aversion parameters *estimated* from the incentivized preference elicitation task without and with controls, respectively. Model (2) and (2c) substitute the inequality aversion parameters with our survey module-based *scores*. The controls include age, immigrant status, income class, highest degree of education, civil status and a dummy for children in the household. The intercept (baseline level) refers to an 18 year old, non-immigrant male with median income and a high school degree who is neither married, nor divorces, separated or widowed, and has no dependent children. p-values: $0 \leq^{***} < 0.01 \leq^{**} < 0.05 \leq^{*} < 0.1$.

All regressions reveal a positive and significant association between behindness aver-

[19] We also observe the expected ordering of support for redistribution across the three preference types discussed earlier, with individuals assigned to the selfish type showing less support compared to those with social preferences. However, this difference is not statistically significant in our data.

sion and volunteering, with slightly stronger effects observed for the survey module-based score. This finding aligns with the intuitive notion that individuals who are more concerned about being left behind are more motivated to engage in volunteering activities.

In Table 13, we analyze the intensive margin of volunteering, focusing on the number of hours spent in volunteering activities per month. The results indicate a positive, though less statistically significant, association between behindness aversion and time investment in volunteering, consistent with expectations. Notably, our survey module-based score demonstrates a stronger ability to detect this relationship compared to the parameters estimated from the incentivized preference elicitation task.

Table 13: Regression Results for Hours Spent in Volunteering `socialHours`

Variable	(1) estimated	(1c) estimated	(2) score	(2c) score
behindness av.	4.309 (2.743)	5.284 (2.831)*	5.794 (2.987)*	6.49 (3.089)**
aheadness av.	-2.524 (2.743)	-3.057 (2.82)	0.731 (3.028)	-0.278 (3.115)
Intercept	2.801 (1.199)**	11.009 (3.081)***	0.447 (1.163)	8.768 (3.118)***
Controls	no	yes	no	yes
R^2	0.005	0.073	0.021	0.085

Note. The response variable are hours per month spent in volunteering activities. Model (1) and (1c) use the inequality aversion parameters *estimated* from the incentivized preference elicitation task without and with controls, respectively. Model (2) and (2c) substitute the inequality aversion parameters with our survey module-based *scores*. The controls include age, immigrant status, income class, highest degree of education, civil status and a dummy for children in the household. The intercept (baseline level) refers to an 18 year old, non-immigrant male with median income and a high school degree who is neither married, nor divorces, separated or widowed, and has no dependent children. p-values: $0 \leq^{***} < 0.01 \leq^{**} < 0.05 \leq^{*} < 0.1$.

Finally, we present the results for the hypothetical donation question proposed by Falk et al. (2023). In this scenario, respondents were asked to imagine winning $1,000 in a lottery and to decide whether (and how much) they would donate to charity. Table 14 reports the results for the extensive margin, focusing on whether respondents would choose to donate. Further analysis on the intensive margin, examining the amount they would donate, are deferred to Appendix A.10.

Our survey module-based score demonstrates significantly stronger predictive power in determining whether respondents would choose to donate. Both the magnitude and statistical significance of the coefficients surpass those observed in models using the estimated preference parameters (Model (1) and (1c)). Consistent with expectations, individuals with higher inequality aversion are more likely to donate after winning the hypothetical lottery.

Overall, our findings highlight the strong external validity of the survey module-based scores across multiple domains. Appendix A.10 provides further evidence by reporting bivariate associations between our preference measures and the survey responses, along with analyses of additional survey questions to extend the robustness of our conclusions.

Table 14: Regression Results for Participation in Giving after Lottery Win `socialHyp1`

Variable	(1) estimated	(1c) estimated	(2) score	(2c) score
behindness av.	0.129 (0.108)	0.061 (0.109)	0.304 (0.114)***	0.271 (0.115)**
aheadness av.	0.119 (0.108)	0.188 (0.109)*	0.258 (0.115)**	0.303 (0.116)***
Intercept	0.402 (0.047)***	0.166 (0.119)	0.25 (0.044)***	-0.003 (0.116)
Controls	no	yes	no	yes
R^2	0.018	0.113	0.092	0.183

Note. The response variable is binary. The reported results are for a linear probability model. Model (1) and (1c) use the inequality aversion parameters *estimated* from the incentivized preference elicitation task without and with controls, respectively. Model (2) and (2c) substitute the inequality aversion parameters with our survey module-based *scores*. The controls include age, immigrant status, income class, highest degree of education, civil status and a dummy for children in the household. The intercept (baseline level) refers to an 18 year old, non-immigrant male with median income and a high school degree who is neither married, nor divorces, separated or widowed, and has no dependent children. p-values: $0 \leq^{***} < 0.01 \leq^{**} < 0.05 \leq^{*} < 0.1$.

6 Conclusion

This study introduces a novel survey module designed to measure inequality aversion and altruism, with an emphasis on simplicity, scalability, and broad applicability. By leveraging data from a representative U.S. population sample, we demonstrate our survey module's capacity to capture behavioral variation in incentivized experiments while maintaining practicality for use in diverse settings. The *"Hearts-and-Minds"* module is specifically crafted to enable applications across a wide range of contexts, from controlled laboratory studies to broad-scale population surveys, especially in scenarios where resources are limited or rapid measurement is required. The survey items can be elicited in just 1-2 minutes, ensuring minimal burden on respondents and making the module highly efficient for broad implementation. The survey module provides reliable and externally valid measures of inequality aversion, making it a valuable tool for examining how individuals respond to unequal resource distributions in various economic and social environments. Using a careful item-selection process, enhanced by machine learning techniques, we ensure that the module is parsimonious yet predictive. This methodology allows researchers to study inequality aversion effectively in diverse contexts without the financial and logistical constraints associated with incentivized experiments. Notably, the module's performance on new data—as tested on a holdout set—demonstrates its generalizability and predictive capabilities.

While some loss in predictive power is observed when comparing module-based scores to real-incentivized preference measures, our external validity exercises reveal that the module-based scores outperform incentivized measures in most cases. This trade-off highlights the strength of the module in broader applicability, particularly for field studies and policy-relevant research. Moreover, our findings underscore the potential of this approach to predict real-world social behaviors, such as altruistic actions and attitudes toward inequality, validating its utility in practical applications. The transparency of our methodology enhances the adaptability of the module for specific research objectives, making it a versatile tool for future investigations. However, since the module has been tested only on a representative U.S. population sample, additional validation is essential across diverse samples, including those from different countries and socioeconomic back-

grounds. Such efforts are crucial to establishing the generalizability of our approach and the robustness of the variable selection process. Expanding data collection to encompass varied populations will also provide valuable input for model training. In particular, it will build a comprehensive database that future versions of the module can leverage to improve predictive accuracy. Future research should prioritize cross-cultural validation and refinement of the module for use in different settings. Additionally, integrating the module with other dimensions of social preferences, such as trust and reciprocity, could offer a more comprehensive understanding of social behaviors. By doing so, we aim to foster deeper insights into the drivers of social preferences and their implications for economic and policy decisions. Overall, this work represents a foundational step toward the development of accessible and generalizable measurement tools for inequality aversion and altruism. Our approach demonstrates the feasibility of scalable survey-based metrics that balance predictive power with practicality.

A Appendix

A.1 Attention Checks

We used three attention checks, also referred to as "screeners," adapted from Berinsky et al. (2021). These asked respondents about the most important problems facing the country, their favorite colors, and news websites. We positioned the screeners so that they were equally spaced throughout the survey. Specifically, `screener1` appeared before the choice tasks, `screener2` after the choice tasks, and `screener3` midway through the survey items. The screeners were presented as follows and in the following order.

Table 15: Attention check items

Item	Description
`screener1`	"*Research shows that questions considered important by some people can influence their opinions on other topics. We also want to know if you are paying attention to the survey. If you do, please ignore the question below and select 'Crime'. Which of the following issues faced by the nation do you think is the most important?*"
`screener2`	"*Some research has shown that individual preferences and knowledge, as well as external factors, can have a significant impact on the decision-making process. To show that you have read carefully, choose 'Pink' from the options below, regardless of your favorite color. Yes, in order to show us that you are paying attention to this survey, please select 'Pink'. What is your favorite color?*"
`screener3`	"*When major news breaks, people often go online to find up-to-the-minute details on current events. We also want to know if you are paying attention to the survey. To show us that you do, please ignore the following question and select 'ABC News' as your answer. When major news breaks, which news website do you visit first?*"

Note. The alternative are as follows. For `screener1`: "*Health care*", "*Unemployment*", "*Public debt*", "*War*", "*Crime*", "*Education*", "*International relations*". For `screener2`: "*White*", "*Black*", "*Red*", "*Pink*", "*Green*", "*Blue*". For `screener3`: "*The New York Times*", "*The Washington Post*", "*CNN*", "*NBC*", "*USA Today*", "*ABC News*", "*CBS News*".

A.2 Representativeness

We targeted a sample of approximately 500 individuals from the U.S. adult population, aiming for representativeness based on three stratification criteria: age group, gender, and ethnicity. The following three tables illustrate that, after excluding participants who failed the three attention checks, the actual proportions in our sample closely align with the target quotas. Deviations per category are generally within ± 1 percentage point, demonstrating that we come very close to the targeted values.

Table 16: Age group

Age group	Target proportion	Actual proportion	Deviation
18 to 24	0.120	0.116	-0.004
25 to 34	0.173	0.172	-0.002
35 to 44	0.169	0.174	0.004
45 to 54	0.159	0.166	0.006
55 to 100	0.378	0.373	-0.005

Table 17: Gender

Gender	Target proportion	Actual proportion	Deviation
Female	0.508	0.499	-0.009
Male	0.492	0.501	0.009

Table 18: Ethnicity

Ethnicity	Target proportion	Actual proportion	Deviation
Asian	0.062	0.070	0.008
Black	0.118	0.116	-0.002
Mixed	0.104	0.116	0.012
Other	0.080	0.076	-0.004
White	0.637	0.623	-0.015

A.3 Survey Items

Table 19: Altruism items

Item	Description
altruismGen1	"*Are you generally willing to share with others without expecting something in return, or are you not willing to do so?*"
altruismContext4	"*Are you generally willing to share with strangers, or are you not willing to do so?*"
altruismDescription1	"*I feel personally responsible for helping others when I am in a position to do so.*"
altruismDescription2	"*I would feel uncomfortable keeping all available resources for myself while others have less.*"
altruismDescription3	"*I value the well-being of others more than maximizing my own personal benefit.*"
altruismDescription4	"*I would rather give to others than see them go without, even if it means I have less.*"
altruismDescription5	"*I believe that sharing with others, even when not required, is the right thing to do.*"
altruismDescription6	"*When I have the chance to give, I do so willingly, regardless of who benefits.*"
altruismDescription7	"*I feel fulfilled when I can give something to others, even if it costs me personally.*"
altruismDescription8	"*I am willing to share what I have with others, whether I know them well or not.*"
altruismDescription9	"*If I had the opportunity to help someone financially, I would, even if it is a complete stranger.*"

Note. The scale is as follows. For `Gen` item: "0: completely unwilling to do so" to "10: very willing to do so." For `Context` item: "0: completely unwilling to share with strangers" to "10: very willing to share with strangers." For `Description` items: "0: does not describe me at all" to "10: describes me perfectly." The items `altruismGen1` and `altruismContext4` are adapted from Falk et al. (2022) and have been rephrased in what we believe to be a simpler and more accessible form.

Table 20: Comparison items

Item	Description
comparisonGen1	*"Do you generally compare what you have with others or not?"*
comparisonContext4	*"Do you generally compare what you have with strangers or not?"*
comparisonDescription1	*"Overall, I am affected by what others have compared to what I have."*
comparisonDescription2	*"Overall, I feel a sense of injustice when others have more than I do."*
comparisonDescription3	*"Overall, I am uneasy when I am better off than others."*
comparisonDescription4	*"Whether others have more or less than I do is irrelevant to me."*
comparisonDescription5	*"It does not affect me if I am better off than someone else."*
comparisonDescription6	*"In a situation where wealth is redistributed, I am satisfied as long as I get something, even if someone else gets much more."*
comparisonDescription7	*"I particularly enjoy situations where I am better off than others."*
comparisonDescription8	*"When I see someone enjoying more resources, I feel a desire to have the same."*
comparisonDescription9	*"I would feel uncomfortable if I perceive advantages or privileges that are not perceived by others."*
comparisonDescription10	*"I feel a sense of injustice when some people have significantly less than what I have."*

Note. The scale is as follows. For `Gen` item: "0: completely unwilling to do so" to "10: very willing to do so." For `Context` item: "0: I absolutely do not compare what I have with strangers" to "10: I absolutely compare what I have with strangers." For `Description` items: "0: does not describe me at all" to "10: describes me perfectly."

Table 21: Inequality aversion items

Item	Description
ineqavGen1	*"Are you generally willing to redistribute resources with others to reduce inequality, or are you not inclined to do so?"*
ineqavContext4	*"Are you generally willing to redistribute resources with strangers to reduce inequality, or are you not inclined to do so?"*
ineqavDescription1	*"I believe it's important to share equally with others, even if I don't know them personally."*
ineqavDescription2	*"I would prioritize equity over maximizing my own benefits if I were in a situation where I had to distribute resources with others."*
ineqavDescription3	*"If I have the choice to distribute resources with strangers, I would rather keep more for myself and give less to others."*
ineqavDescription4	*"If I have the choice to distribute resources with strangers, I would rather give more to others and keep less for myself."*
ineqavDescription5	*"When I have more than someone else, I feel like I should share what I have."*
ineqavDescription6	*"In situations where I would earn more than others for the same effort, I would feel the need to limit my income at a certain point, even if I could earn more."*
ineqavDescription7	*"I would be willing to sacrifice a large part of my income to slightly reduce that of those less well off than me."*
ineqavDescription8	*"In situations where others would earn more than me for the same effort, I would be willing to set an income limit for everyone."*
ineqavDescription9	*"I would be willing to sacrifice a little of my income to drastically reduce that of the most fortunate."*

Note. The scale is as follows. For `Gen` item: "0: completely unwilling to do so" to "10: very willing to do so." For `Context` item: "0: completely unwilling to redistribute resources with strangers to reduce inequality" to "10: very willing to redistribute resources with strangers to reduce inequality." For `Description` items: "0: does not describe me at all" to "10: describes me perfectly." We intentionally reverse-coded `ineqavDescription3` to check participants' consistency in responses, although this item is not intended to serve as a screener. We do observe consistency in responses, as the α and β parameter values are positively correlated when the scale is adjusted (see Figure 6).

A.4 Hypothetical Questions

The first question (`strategyHN`) is a hypothetical version of the incentivized choice tasks, involving a trade-off between the self's payoff and the other's payoff. The other questions (`socialHyp1` and `socialHyp2`) are adapted from Falk et al. (2023), but decomposed into two parts: the subject first indicates whether she/he would donate to charity, and only then specifies the amount (we believe this slight modification reduces priming).

Table 22: Hypothetical questions

Item	Description
`strategyHN`	"*Imagine you are in a situation where you have to distribute money between yourself and an anonymous person. Neither of you will ever see or interact with the other. You have absolutely no information about the other person's circumstances (such as his/her wealth). The only thing you know is that nobody, except you and the other person, will ever know your choice. What would you do? I would...*"
`socialHyp1`	"*Imagine the following situation: you won $1,000 in a lottery. Considering your current situation, would you donate a part of your gains to charity?*"
`socialHyp2`	"*If you would, how much would you donate to charity? (Please indicate '0' if you would not.)*"

Note. The alternatives are as follows (with the associated strategy in parentheses). For `strategyHN`: "*keep everything for myself*" (`selfish`), "*take a larger portion for myself and leave a smaller portion for the other*" (`ineqselfish`), "*make an approximately equal distribution between myself and the other person*" (`egalitarian`), "*take a smaller portion for myself and leave a larger portion to the other person*" (`ineqaltruism`), "*give everything to the other person*" (`altruism`), "*do something else (see below)*" (`other`: open text field). For `socialHyp1`: "*Yes/No*". For `socialHyp2`: open text field.

Table 23 documents the number of respondents that chose one of the six possible strategies in the hypothetical survey question. 187 respondents (37.3%) stated the `selfish` or the mainly selfish (`ineqselfish`) strategy. 301 respondents (60%) stated the `egalitarian` strategy. Only a few subjects chose one of the other strategies.

Table 23: Number of respondents' strategies in the hypoethetical question `strategyHN`

Variable	Count
`selfish`	72
`ineqselfish`	115
`egalitarian`	301
`altruism`	1
`ineqaltruism`	4
`other`	9

A.5 Real-World Behavior

We adapted the real-world behavior questions from Falk et al. (2023) by replacing references to "charity" with "association/volunteering community" to make them more general, except for one question, which specifically addressed donations. We also included two items assessing people's general approval or disapproval of inequality in the U.S. and their support for policies aimed at reducing inequality.

Table 24: Volunteering and Social Responsibility Items

Item	Description
`socialMember1`	"*I am a member of an association/volunteering community.*"
`socialHours`	"*Please specify as precisely as possible how many hours per month you volunteer for an association/volunteering community. (If you do not, simply indicate '0'.)*"
`socialOthers`	"*How many people (approximately) know that you commit time to an association/volunteering community? (If you do not, simply indicate '0'.)*"
`socialMember2`	"*I am a donor to an association/volunteering community (regular or not).*"
`socialAmount`	"*Please specify as precisely as possible what amount you have given to charity over the past year. (If you have not, please enter '0'.)*"
`ineqPolicy`	"*I support policies aimed at reducing inequality, such as taxing the rich to help the poor.*"

Note. The alternatives are as follows. For `socialMember1` and `socialMember2`: "*Yes/No*". For `socialHours`, `socialOthers` and `socialAmount`: open text field. For `ineqPolicy`: "*0: does not describe me at all*" to "*10: describes me perfectly*".

A.6 Type Characterization: Results for Two and Four Types

Table 25 and 26 show the proportions of subjects assigned to the emerging types. The Alluvial plot in Figure 12 depicts how subjects transition between assigned types when enforcing two, three, four and five types. As argued in the main text, the three type clustering yields a clear interpretation of the types. However, it appears that parts of this interpretation gets lost when forcing the algorithm to return only two types. In the 2-type clustering, the first type (Type 1) is an amalgam of selfish (red for three types) and altruistic (green for three types). The second type (Type 2) of the 2-type clustering contains nearly all inequality averse subjects from the three type clustering, but also a substantial portion of the altruists that we found there. Similarly, going from three to four and more types yields smaller types with less clear interpretation.

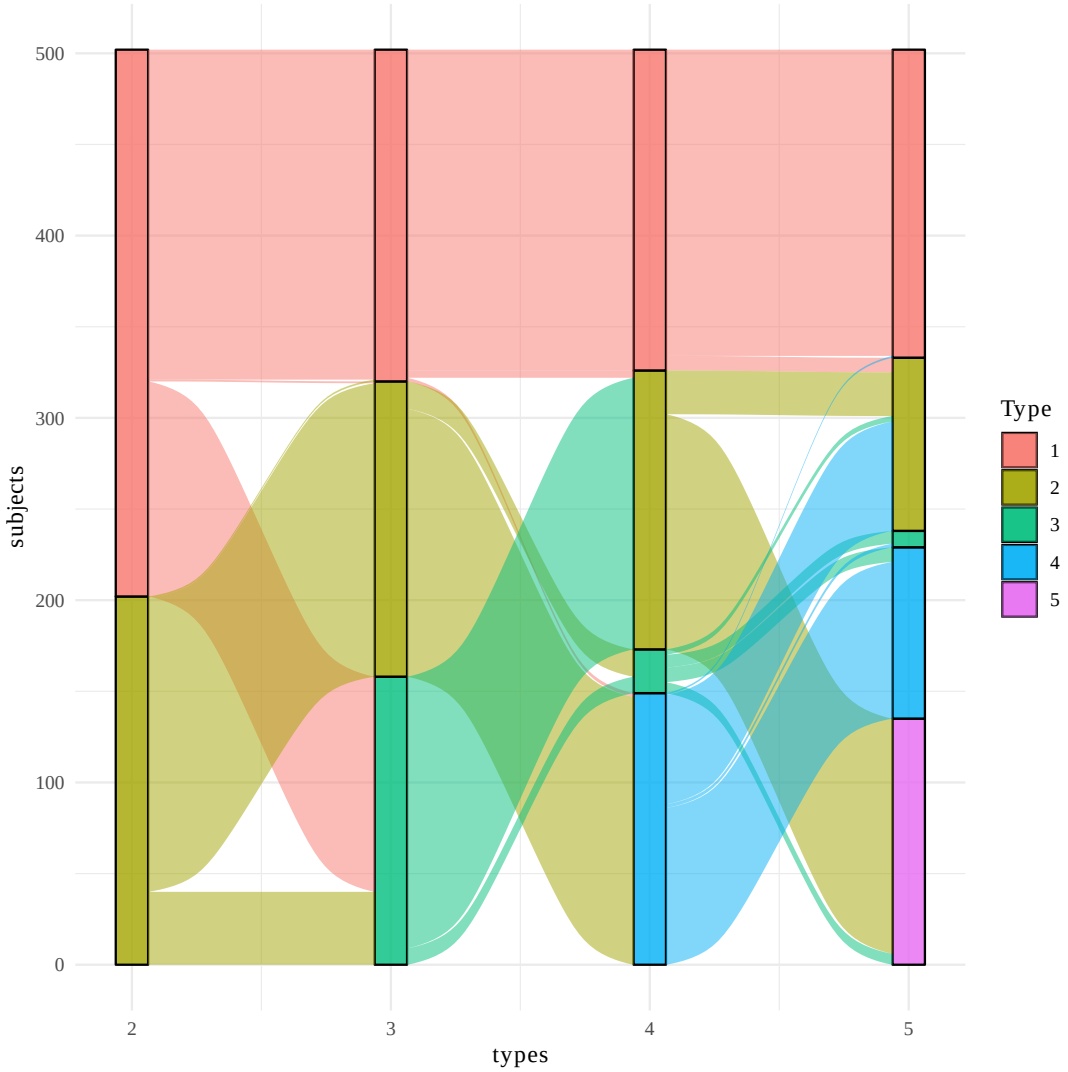
Table 25: Distribution of preference types | $k = 2$

Type	Proportion
1	59.76%
2	40.24%

Table 26: Distribution of preference types | $k = 4$

Type	Proportion
1	35.06%
2	30.48%
3	4.78%
4	29.68%

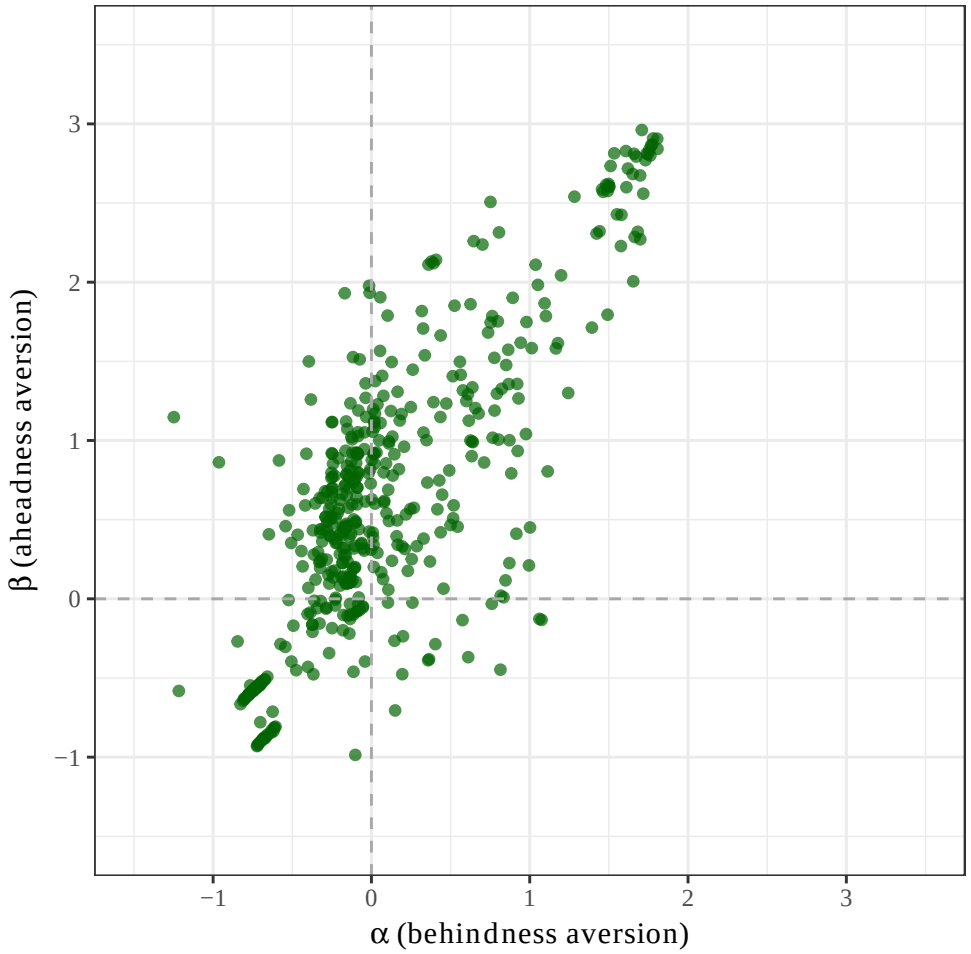
Figure 12: Alluvial plot



A.7 Structural Estimation Results

Figure 13 depicts the association between individual aheadness and behindness aversion parameters. The positive correlation between domain-specific inequality aversion discussed in the main text is clearly visible.

Figure 13: Association between aheadness and behindness aversion parameters



A.8 Structural Estimation Results by Type

Figure 14 illustrates the distribution of aheadness (β_i) and behindness (α_i) aversion parameters conditional on type assignment. The interpretation of the types becomes immediately apparent. Type 1 is best described by an average β_i close to zero, but a slightly negative α_i. In other words, this type is selfish and even a bit spiteful when being behind. Type 2 is characterized by largely positive inequality aversion in both the aheadness and the behindness domain. We therefore label this type as inequality averse. Type 3 exhibits a more asymmetric behavior with mostly positive β_i (aheadness aversion), but α_i's close to zero (selfishness in the behindness domain). Consistent with this, we coin this type "altruistic".

Figure 14: Within-type distribution of aheadness and behindness aversion parameters

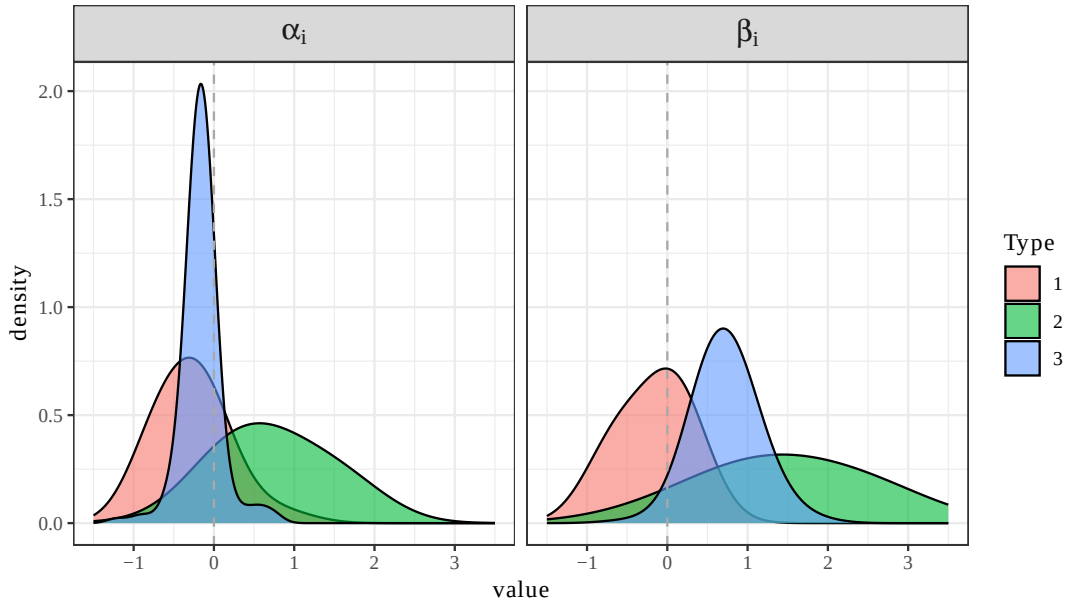
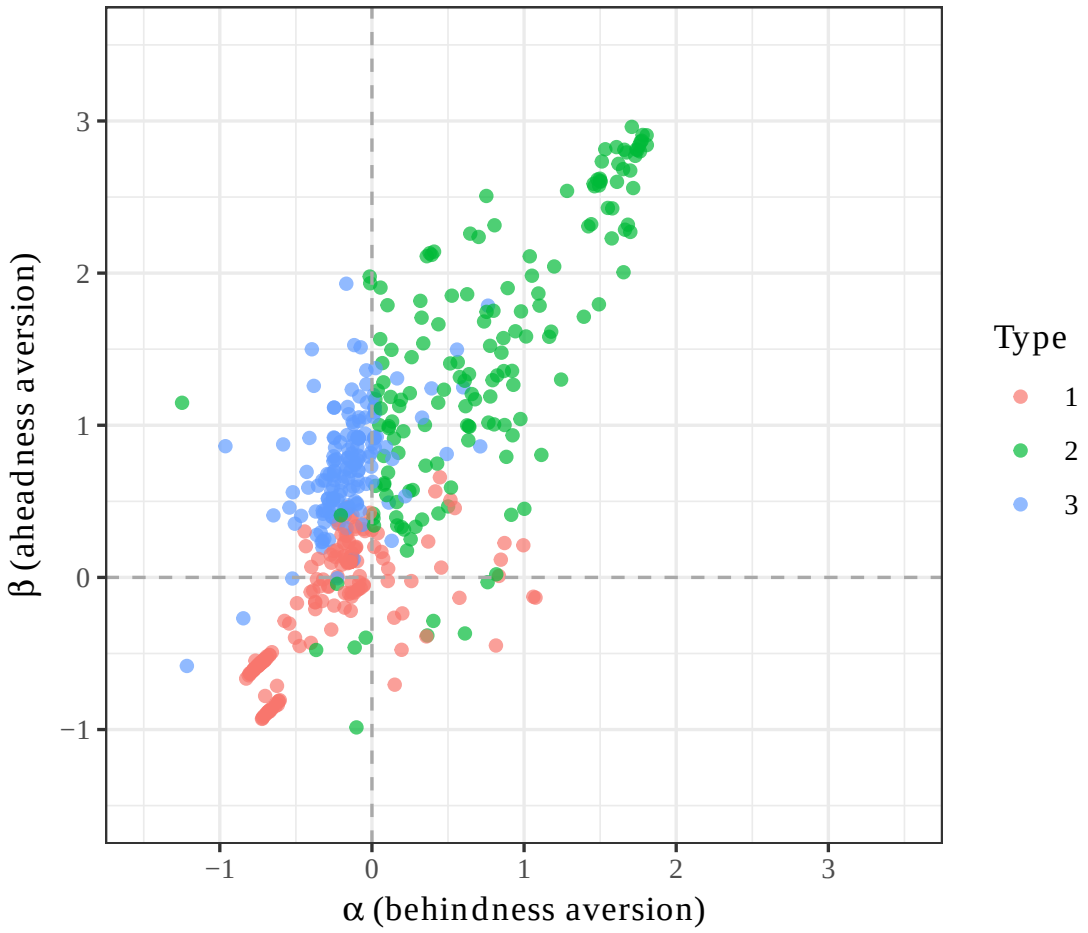


Figure 15 makes these results visible in the scatter plot. In fact, the selfish type's parameters scatter around zero. The inequality averse type shows a more heterogeneous distribution with largely positive inequality aversion in the aheadness and behindness domain. Lastly, the altruistic type's parameters lie mostly in the upper left quadrant of the figure.

Figure 15: Association between aheadness and behindness aversion parameters by type



A.9 Ability of the Structural Model to Capture Features of the Data

Figures 16 and 17 split the α_i and β_i parameters into deciles labeled as D1 (low value) to D10 (high value). As the figures illustrate, subjects who got estimated a high value of the parameters indeed exhibit more inequality aversion in the respective domain. Thereby, α_i seems to more clearly separate the deciles in the behindness domain, whereas β_i seems to more clearly separate the deciles in the aheadness domain. Note, however, that the two parameters are highly correlated in our data.

Figure 16: Deciles α_i (behindness aversion)

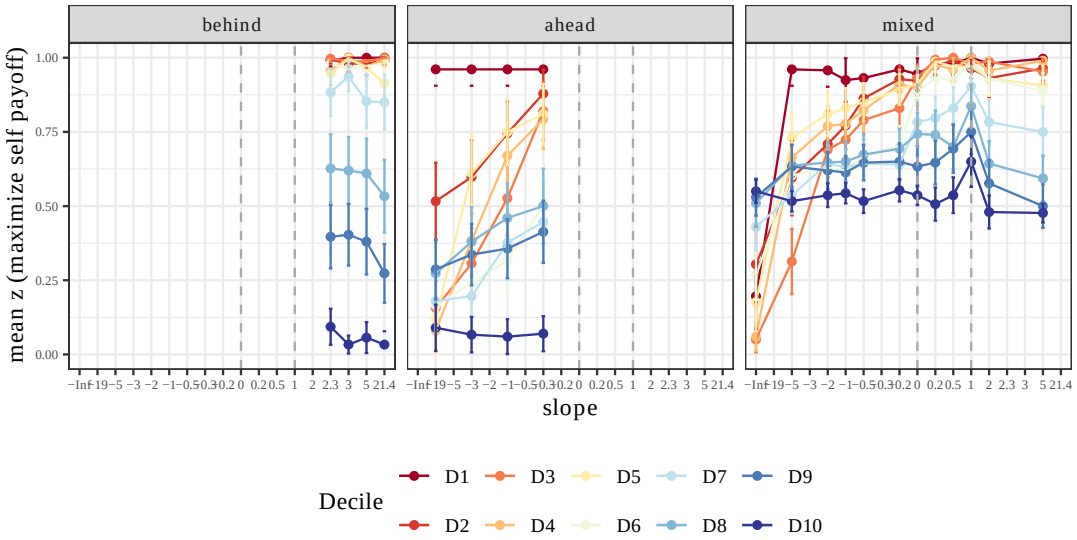
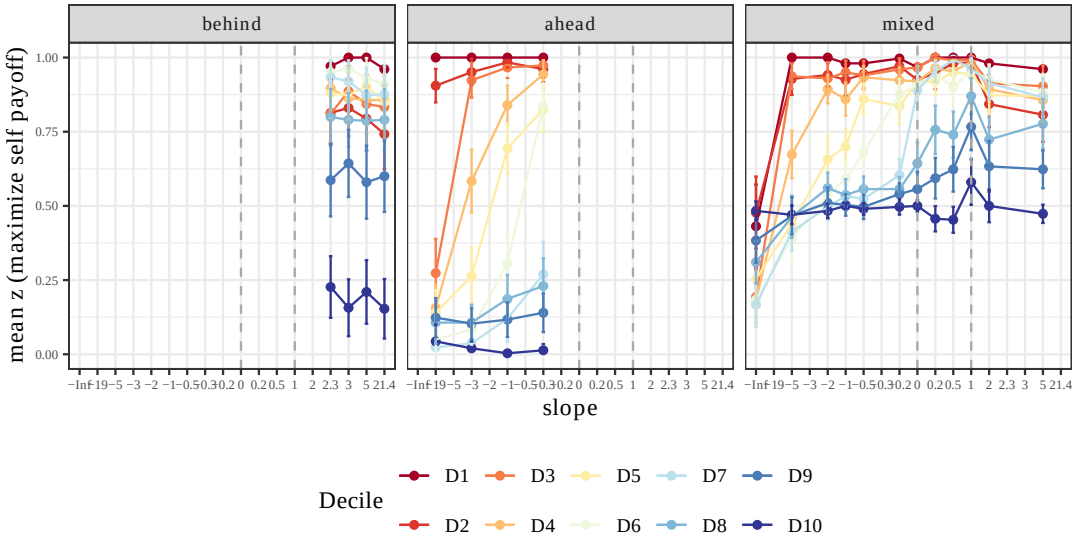


Figure 17: Deciles β_i (aheadness aversion)



A.10 External Validity

Tables 27 and 28 document the bivariate relationships between our real-world and hypothetical survey items, and the estimated inequality aversion parameters and our module-based scores.

Table 27 presents results for Spearman rank correlation tests on the association between the continuous variables and the preference measures. Our module-based score is more strongly and significantly associated with the different stated behaviors than the preference parameters obtained from estimation.

Table 27: Bivariate associations between estimated preference parameters/module indices and continuous real-world behaviors

Variable	(i) behind estim.	(ii) behind score	(iii) ahead estim.	(iv) ahead score
`socialHours`	0.127 (0.004)	0.238 (0.000)	0.069 (0.124)	0.198 (0.000)
`socialOthers`	0.105 (0.018)	0.210 (0.000)	0.081 (0.070)	0.189 (0.000)
`socialAmount`	0.064 (0.152)	0.128 (0.004)	0.061 (0.169)	0.131 (0.003)
`socialHyp2`	0.191 (0.000)	0.364 (0.000)	0.183 (0.000)	0.330 (0.000)
`ineqPolicy`	0.055 (0.223)	0.151 (0.001)	0.125 (0.005)	0.237 (0.000)

Note. `behind` and ahead refer to the behindness aversion (α_i or score) and aheadness aversion (β_i or score) parameters, respectively. The table reports Spearman rank correlations between preference parameters obtained from the incentivized elicitation task, α_i and β_i (see columns (i) and (iii)), and a series of self-stated behaviors. It reports the same for our behindness aversion index prediction (column (iii)) and our aheadness aversion index prediction (column (iv)). p-values are stated in parentheses.

Table 28 reports results of Mann-Whitney U tests. More specifically, we test whether inequality aversion is higher for those who are participating in volunteering and donate to charities (one-sided test). As we see, this is indeed the case for all variables, with our scores performing better than estimated parameters.

Table 28: Bivariate associations between estimated preference parameters / module indices and binary real-world behaviors

Variable	(i) behind estim.	(ii) behind score	(iii) ahead estim.	(iv) ahead score
`socialMember1`	0.180 (0.001)	0.138 (0.000)	0.158 (0.037)	0.172 (0.000)
`socialMember2`	0.056 (0.098)	0.075 (0.001)	0.027 (0.232)	0.139 (0.000)
`socialHyp1`	0.133 (0.003)	0.139 (0.000)	0.187 (0.003)	0.229 (0.000)

Note. `behind` and ahead refer to the behindness aversion (α_i or score) and aheadness aversion (β_i or score) parameters, respectively. The table reports differences in the means of the parameters for Variable=1 - Variable=0. The p-values are for one-sided Mann-Whitney U tests.

Table 29 presents regression results on the intensive margin of charitable giving after a hypothetical lottery win. The results here are a bit less clear, but according to our scores, there is evidence that behindness aversion is associated positively with the donated amount.

Similar results emerge for monetary donations to a volunteering community (see Table 30). Here it is the aheadness aversion that is positively associated with donations. Once again, it is our score that picks up this association, while estimated parameters do

Table 29: Regression Results for Amount of Donations after Lottery Win `socialHyp2`

Variable	(1) estimated	(1c) estimated	(2) score	(2c) score
behindness av.	116.176 (78.104)	102.922 (80.552)	190.45 (84.977) **	164.518 (87.841) *
aheadness av.	120.767 (78.104)	142.933 (80.251) *	100.69 (86.169)	126.847 (88.6)
Controls	no	yes	no	yes
R^2	0.03	0.096	0.047	0.108

Note. The response variable is the amount donated after a hypothetical lottery win. Model (1) and (1c) use the inequality aversion parameters *estimated* from the incentivized preference elicitation task without and with controls, respectively. Model (2) and (2c) substitute the inequality aversion parameters with our survey module-based *scores*. The controls include age, immigrant status, income class, highest degree of education, civil status and a dummy for children in the household. p-values: $0 \leq^{***} < 0.01 \leq^{**} < 0.05 \leq^{*} < 0.1$.

not.

Table 30: Regression Results for Donations to Volunteering Community `socialMember2`

Variable	(1) estimated	(1c) estimated	(2) score	(2c) score
behind	0.114 (0.104)	0.07 (0.105)	0.078 (0.113)	0.023 (0.114)
ahead	-0.025 (0.104)	0.022 (0.105)	0.214 (0.115) *	0.271 (0.115) **
Controls	no	yes	no	yes
R^2	0.003	0.11	0.027	0.134

Note. The response variable is a binary variable for whether the respondent donates to a volunteering community. We estimate a linear probability model. Model (1) and (1c) use the inequality aversion parameters *estimated* from the incentivized preference elicitation task without and with controls, respectively. Model (2) and (2c) substitute the inequality aversion parameters with our survey module-based *scores*. The controls include age, immigrant status, income class, highest degree of education, civil status and a dummy for children in the household. p-values: $0 \leq^{***} < 0.01 \leq^{**} < 0.05 \leq^{*} < 0.1$.

Tables 31 and 32 show two instances where we fail to detect any association with aheadness and behindness aversion. The results are consistent between estimated parameters and scores. While we find clear bivariate associations between these variables and our score (see Table 27), we do not find any support for such associations in the regressions. This results is not particularly surprising, however. The `socialOthers` item should possibly only be weakly related to own preferences. The number of people knowing about respondents' volunteering activities may crucially depend on other factors (social network and nature of the association, etc.), factors we cannot control for. Similarly, the amount donated to charities (`socialAmount`) is heavily influenced by wealth and income. We have a rough measure for the latter and control for it in the regressions. However, there are likely more complex interactions at play here (see Epper et al. (2024) who use third-party registered data on charitable donations for a discussion).

Table 31: Regression Results for People Knowing about Volunteering `socialOthers`

Variable	(1) estimated	(1c) estimated	(2) score	(2c) score
behind	5.012 (9.514)	4.644 (9.812)	7.982 (10.414)	7.166 (10.762)
ahead	1.306 (9.514)	-0.972 (9.775)	5.817 (10.56)	1.424 (10.855)
Controls	no	yes	no	yes
R^2	0.002	0.072	0.007	0.074

Note. The response variable is the number of people the respondent knows that she/he commit time in volunteering. Model (1) and (1c) use the inequality aversion parameters *estimated* from the incentivized preference elicitation task without and with controls, respectively. Model (2) and (2c) substitute the inequality aversion parameters with our survey module-based *scores*. The controls include age, immigrant status, income class, highest degree of education, civil status and a dummy for children in the household. p-values: $0 \leq^{***} < 0.01 \leq^{**} < 0.05 \leq^{*} < 0.1$.

Table 32: Regression Results for Donations to Charities `socialAmount`

Variable	(1) estimated	(1c) estimated	(2) score	(2c) score
behind	2329.21 (1960.58)	2021.82 (1921.78)	-79.02 (2150.58)	507.08 (2110.67)
ahead	-2361.18 (1960.58)	-2074.41 (1914.59)	2247.68 (2180.74)	1061.82 (2128.91)
Controls	no	yes	no	yes
R^2	0.003	0.163	0.005	0.163

Note. The response variable is the (self-reported) amount of U.S. dollars donated to charities over the past year. Model (1) and (1c) use the inequality aversion parameters *estimated* from the incentivized preference elicitation task without and with controls, respectively. Model (2) and (2c) substitute the inequality aversion parameters with our survey module-based *scores*. The controls include age, immigrant status, income class, highest degree of education, civil status and a dummy for children in the household. p-values: $0 \leq^{***} < 0.01 \leq^{**} < 0.05 \leq^{*} < 0.1$.

References

Alesina, A. and G.-M. Angeletos (2005). Fairness and redistribution. *American Economic Review 95*(4), 960–980.

Alesina, A. and P. Giuliano (2011). Preferences for redistribution. In J. Benhabib, A. Bisin, and M. Jackson (Eds.), *Handbook of Social Economics, Vol. 1A*, pp. 93–131. Amsterdam: North-Holland.

Baillon, A., H. Bleichrodt, and V. Spinu (2019). Searching for the reference point. *Management Science 66*(1), 93–112.

Berinsky, A. J., M. F. Margolis, M. W. Sances, and C. Warshaw (2021). Using screeners to measure respondent attention on self-administered surveys: Which items and how many? *Political Science Research and Methods 9*(2), 430–437.

Chen, T. and C. Guestrin (2016). Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Corneo, G. and H. P. Grüner (2002). Individual preferences for political redistribution. *Journal of Public Economics 83*(1), 83–107.

Cowell, F. A. (2011). *Measuring Inequality* (3rd ed.). Oxford, UK: Oxford University Press.

Decancq, K., M. Fleurbaey, and F. Maniquet (2019). Multidimensional poverty measurement with individual preferences. *Journal of Economic Inequality 17*(1), 29–49.

Decancq, K., M. Fleurbaey, and E. Schokkaert (2017). Wellbeing inequality and preference heterogeneity. *Economica 84*(334), 210–238.

Epper, T., E. Fehr, H. Fehr-Duda, C. T. Kreiner, D. D. Lassen, S. Leth-Petersen, and G. N. Rasmussen (2020). Time discounting and wealth inequality. *American Economic Review 110*(4), 1177–1205.

Epper, T. F., E. Fehr, C. T. Kreiner, S. Leth-Petersen, I. S. Olufsen, and P. E. Skov (2024). Inequality aversion predicts support for public and private redistribution. *Proceedings of the National Academy of Sciences 121*(39), e2401445121.

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

Falk, A., A. Becker, T. Dohmen, D. Huffman, and U. Sunde (2023). The preference survey module: a validated instrument for measuring risk, time, and social preferences. *Management Science 69*(4), 1935–1950.

Fehr, E. and G. Charness (2025). Social preferences: fundamental characteristics and economic consequences. *Journal of Economic Literature (forthcoming)*.

Fehr, E., T. Epper, and J. Senn (2023). The fundamental properties, stability and predictive power of distributional preferences. *mimeo*, 1–56.

Fehr, E., T. Epper, and J. Senn (2024). Social preferences and redistributive politics. *The Review of Economics and Statistics*, 1–45.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics 114*(3), 817–868.

Fleurbaey, M. and S. Zuber (2024). Unequal inequality aversion within and among countries and generations. *Journal of Economic Inequality*.

Fong, C. (2001). Social preferences, self-interest, and the demand for redistribution. *Journal of Public Economics 82*(2), 225–246.

Guillaud, E. (2013). Preferences for redistribution: an empirical analysis over 33 countries. *Journal of Economic Inequality 11*(1), 57–78.

Hvidberg, K. B., C. T. Kreiner, and S. Stantcheva (2023). Social positions and fairness views on inequality. *The Review of Economic Studies 90*(6), 3083–3118.

Kulis, B. and M. I. Jordan (2012). Revisiting k-means: New algorithms via Bayesian nonparametrics. *Proceedings of the 29th International Conference of Machine Learning*.

Lecouteux, G. and I. Mitrouchev (2024). The view from manywhere: normative economics with context-dependent preferences. *Economics and Philosophy 40*(2), 374–396.

McFadden, D. (1981). Econometric models of probabilistic choice. In C. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 198–272. MIT Press.

Mengel, F. and E. Weidenholzer (2023). Preferences for redistribution. *Journal of Economic Surveys 37*(5), 1660–1677.

Piketty, T. and E. Saez (2003). Income inequality in the United States, 1913–1998. *The Quarterly Journal of Economics 118*(1), 1–39.

Piketty, T. and E. Saez (2014). Inequality in the long run. *Science 344*(6186), 838–843.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn and A. W. Tucker (Eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton: Princeton University Press.

Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. *Science 211*(4481), 453–458.