iRisk WORKING PAPER SERIES

The Fundamental Properties, Stability and Predictive Power of Distributional Preferences

Ernst Fehr
Department of Economics, Zurich University. Blümlisalpstrasse 10, 8006 Zurich, Switzerland
(ernst.fehr@econ.uzh.ch)

Thomas Epper
IESEG School of Management, Univ. Lille, CNRS, UMR 9221 - LEM - Lille Economie Management, F-59000 Lille, France; iRisk Center on Risk and Uncertainty (thomas.epper@cnrs.fr)

Julien Senn
Department of Economics, Zurich University. Blümlisalpstrasse 10, 8006 Zurich, Switzerland (julien.senn@econ.uzh.ch)

The Fundamental Properties, Stability and Predictive Power of Distributional Preferences*

Ernst Fehr ⓡ Thomas Epper ⓡ Julien Senn

October 10, 2023

Abstract

Parsimony is a desirable feature of economic models but almost all human behaviors are characterized by vast individual variation that appears to defy parsimony. How much parsimony do we need to give up to capture the fundamental aspects of a population's distributional preferences and to maintain high predictive ability? Using a Bayesian nonparametric clustering method that makes the trade-off between parsimony and descriptive accuracy explicit, we show that three preference types—an inequality averse, an altruistic and a predominantly selfish type—capture the essence of behavioral heterogeneity. These types independently emerge in four different data sets and are strikingly stable over time. They predict out-of-sample behavior equally well as a model that permits all individuals to differ and substantially better than a representative agent model and a state-of-the-art machine learning algorithm. Thus, a parsimonious model with three stable types captures key characteristics of distributional preferences and has excellent predictive power.

Key Words: Distributional Preferences, Altruism, Inequality Aversion, Preference Heterogeneity, Stability, Out-of-Sample Prediction, Parsimony, Bayesian Nonparametrics.

JEL Codes: D31, D63, C49, C90

1 Introduction

A large and growing body of research suggests that social preferences play an important role in many economic and social domains.[1] It is thus crucial to understand the properties and the distribution of these preferences in the broader population, and to capture the prevailing preference heterogeneity in a parsimonious manner. Parsimony is particularly important in applied contexts, where the degree of complexity that theories can afford (at the individual level) is limited by tractability constraints.

When modelling preferences, the most parsimonious approach is to assume that a representative agent captures the population's preferences. This is, however, particularly problematic in the domain of social and distributional preferences because even a minority of individuals, who would have little weight in the representative agent's preferences, may have a disproportionate influence on equilibrium outcomes. This follows from the fact that individuals with social preferences often display behaviors that change others' incentives, i.e., even the incentives of those without these preferences.[2] Thus, completely neglecting preference heterogeneity may induce seriously misleading conclusions and predictions.

But how much parsimony must we sacrifice to capture the fundamental characteristics of behavioral heterogeneity? In other words, at which level of parsimony do we still capture the key characteristics of behavioral heterogeneity, and when do we start to neglect important aspects? How much predictive power—in terms of the precision of out-of-sample predictions—do we have to give up if we want to remain parsimonious? Finally, and perhaps most importantly, does a stable core of

---

[1]For the role of social preferences and fairness concerns in bilateral bargaining see, e. g., Camerer and Thaler (1995), Camerer and Loewenstein (1993), and Camerer (2011). For their role in labor and goods markets see, e. g., Fehr et al. (1993), Charness (2000), Charness (2000), Bellemare and Shearer (2007), Dur (2009), Gächter and Thöni (2010), Gächter et al. (2013), Kube et al. (2012), Breza et al. (2018). For their role in political economy, collective action and cooperation, see, e. g., Gächter and Thöni (2005), Tyran and Sausgruber (2006), Durante et al. (2014), Kerschbamer and Müller (2020), Fehr ⓡ al. (2021a),Breza et al. (2021), and Breza et al. (2019). For their role in contract design, mechanism design, and institutions see, e. g., Bierbrauer and Netzer (2016), Bierbrauer et al. (2017), Schmidt and Ockenfels (2021), and Fehr et al. (2021b).

[2]A selfish proposer in the ultimatum game, for example, may have a reason to make fair offers even if only a (significant) minority of the responders rejects unfair offers. Likewise, a selfish employer in a gift exchange game may have a reason to pay high, nonmarket clearing wages, although "only" a minority of employees reciprocates too high wages with higher effort. In public good situations, a minority of players willing to punish free-riders can induce selfish players to contribute (see, e.g., Fehr and Schmidt, 1999).

behavioral heterogeneity exist across data sets and time, or are preferences too fluid and shaped by flimsy, uncontrollable details that prevent the identification of stable heterogeneity?

In this paper, we use a Bayesian nonparametric clustering algorithm—the Dirichlet Process Means (DP-means) algorithm (Kulis and Jordan, 2012)—to answer these questions in the context of distributional preferences. A key feature of this method is that it makes the trade-off between parsimony and descriptive accuracy explicit. The algorithm requires the researcher to specify a desired level of precision with which individuals are assigned to different behavioral clusters in terms of the individuals' maximum allowed deviation from the center of their nearest cluster (also called "centroid").[3] Naturally, as more precision is demanded, the description of behavioral heterogeneity becomes richer because more clusters emerge and the behavioral variation within clusters declines. However, this increased precision has a cost in terms of parsimony (since more clusters emerge)—giving rise to the precision (accuracy)–parsimony trade-off.

The DP-means algorithm also has several other attractive properties. Perhaps most importantly, it can be applied to the raw choice data such that it does not require any ex-ante assumptions about the structure of behaviors or preferences. The algorithm enables the identification of behavioral clusters without assumptions on the *number* of existing preference clusters and the behavioral properties (e.g., the utility functions) of the different clusters. Once the level of precision is fixed, the algorithm endogenously determines the number of clusters that minimizes the sum of individuals' deviations from their clusters.

We apply the DP-means algorithm to four different data sets on distributional preferences collected in 2017 and 2020, covering in total 1731 subjects who are broadly representative of the French and the German language populations in Switzerland. The different data sets enable us to examine the stability of the fundamental characteristics of behavioral heterogeneity across data sets with different individuals, as well as across a three-year time period for the same individuals.

We elicited distributional preferences using a variety of incentivized money al-

---

[3]The individuals' deviation from their clusters is measured in terms of the squared Euclidean distance.

location tasks in which the decision maker has to decide how to allocate money between him/herself and another anonymous participant. We use a combination of choice situations in which the decision maker can pay to *increase* the other participant's payoff, *and* situations in which she can pay to *decrease* the other participant's payoff. Thus, our design goes beyond traditional dictator games which elicit social preferences by only allowing individuals to increase the payoff of others at a cost to themselves. We also systematically vary the decision-maker's cost of increasing or decreasing the other's payoff, which gives us the opportunity to identify a wide variety of distributional preferences such as altruism (as in Andreoni and Miller, 2002 or Fisman et al., 2007, 2017), concerns for the total payoff (Charness and Rabin, 2002), envy (Bolton, 1991; Kirchsteiger, 1994), or inequality aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), which all rest on different assumptions about individuals' willingness to pay to increase and/or decrease others' payoffs. The variation in costs provides valuable information about the trade-offs subjects make when increasing or decreasing others' payoffs, information that we can use when we examine the quality of models with different degrees of parsimony in terms of the accuracy of their out-of-sample predictions.

How much parsimony do we have to sacrifice to capture the fundamental and decisive characteristics of behavioral heterogeneity? Do we have to go far in the direction of individual differences in preferences, or could a small number of clusters already capture the essence of the existing heterogeneity? Our results show that one does not have to move far beyond the representative agent model to capture the essence of behavioral heterogeneity. Specifically, we find that *three clusters* with a clear and unambiguous behavioral interpretation describe the essence of the existing behavioral heterogeneity in all four data sets. Moreover, the *same* three clusters—in terms of their behavioral interpretation and in terms of their relative quantitative importance—emerge in each of the four data sets:

1. The biggest cluster always consists of individuals who predominantly make payoff-equalizing choices. These individuals display a general willingness to pay to increase the payoff of others who are worse off, *as well as* a willingness to pay to decrease the payoff of those who are better off. This behavior is consistent with *inequality aversion* as modeled in Fehr and Schmidt (1999) or Bolton and

Ockenfels (2000).

2. The second somewhat smaller but still large cluster always consists of individuals who display a fundamentally different kind of other-regarding behavior. They are unwilling to reduce inequality by decreasing the payoff of those who are better off but, similar to the inequality averse individuals, they are generally willing to pay to increase the payoff of those who are worse off. Their behavior is thus consistent with *altruistic preferences* as modeled in Charness and Rabin (2002) or Fisman et al. (2007, 2017), for example.

3. The smallest, yet still substantial, cluster consists of individuals who predominantly maximize their own payoff without paying much attention to other individuals' payoffs. We therefore label them as *predominantly selfish*.

The fact that clusters with the same qualitative behavioral properties emerge in all four data sets suggests that the behavioral heterogeneity at the three-cluster level is rather stable across time and data sets. In all samples, the cluster of inequality averse individuals comprises between 45 and 53 percent of the population; the cluster of altruistic individuals comprises between 30 and 42 percent of the population, and the cluster of predominantly selfish individuals comprises between 10 and 24 percent of the population. These findings suggest a relatively stable structure of behavioral heterogeneity at the type-level.

Since the clustering results depend on the desired level of precision, we also ask what happens if we demand more or less precision. We find that demanding less precision such that only two clusters emerge i) dramatically undermines the behavioral interpretation of the clusters because the algorithm systematically pools incompatible behavioral types (i.e., merges them into one cluster), and ii) substantially erodes the stability of the behavioral interpretation of the clusters across data sets. For example, inequality averse and selfish individuals are merged into one cluster in one dataset, while the altruistic and the selfish individuals are merged into one cluster in another dataset. These results suggest that requiring the model to be more parsimonious than three clusters is unsatisfactory as it causes us to neglect important parts of behavioral heterogeneity.

In contrast, when we allow the model to be less parsimonious (by requiring more precision such that a larger number of clusters emerge), we find that no new meaningful and empirically relevant behavioral types emerge. Indeed, when the number of clusters increases to four or five, the newly emerging clusters are populated by an exceedingly small share of individuals (less than 2%) in three of the four data sets, and their behavioral patterns are difficult to interpret. Thus, becoming less parsimonious does not bring fundamentally new insights in terms of novel and empirically relevant behavioral types.

Note that ex-ante it is far from obvious that the population can be characterized by the three types we identified. In principle, it would have been possible to find three other types (e.g., strong altruists, moderate altruists, and behindness averse individuals such as in the student sample of Bruhin et al., 2018), or that the population might be parsimoniously characterized by four different types, etc. However, the fact that we identify the same three types in all four data sets generates confidence in the stability of our three-type distribution.

The superiority of three clusters as a description of the essential aspects of behavioral heterogeneity can also be expressed in terms of the *precision-parsimony frontier*. This frontier is defined as the *smallest* mean squared Euclidean distance of individuals from their clusters for each number of clusters (i.e., each given level of parsimony). In other words, the frontier gives us the highest precision in individuals' assignment to clusters for given levels of parsimony. We computed this frontier for each of the four data sets and find that it exhibits strongly decreasing benefits (measured in terms of increases in precision) of relaxing parsimony. This suggests that it may not be necessary to sacrifice a lot of parsimony to increase precision. Moreover, the frontier has a clear, salient kink at three clusters: when moving from the representative agent model (one cluster) to two or three clusters, rather large precision gains accrue, while the precision gains above three clusters are relatively small. Moving below three clusters is therefore associated with a large loss in precision, while moving above three clusters yields only a small gain in precision.

Thus, considerations based on the precision-parsimony frontier also suggest that three clusters provide the best description of the prevailing behavioral heterogeneity while simultaneously maintaining parsimony. But how much predictive power—in

terms of the accuracy of out-of-sample predictions—are we sacrificing by committing to a relatively parsimonious three-cluster approach? And how much additional predictive power do we gain by moving from a representative agent model to a three-cluster model? To answer these questions, we compare the predictive accuracy of our model with three distributional preference types with the accuracy of (i) the representative agent model and (ii) a model based on each individuals' preferences.

To make *quantitative* out-of-sample predictions, we estimate a distributional two-parameter utility function[4] for (i) the representative agent, (ii) each of the three behavioral types in the three-cluster approach, and (iii) for each individual in our four data sets. We estimate these utility functions on the basis of subjects' decisions for a given set of choice situations (the "estimation set"). We then use these estimated parameters of the utility functions to make point predictions for other choice situations (the "prediction set") for which the subjects also made decisions.[5]

We find that despite the huge increase in the complexity of the individual-level model (2N parameters) relative to the three-type model (6 parameters only), the latter predicts equally well. Indeed, the out-of-sample hit rate[6] of the three-type model varies between 66.5 percent and 72.5 percent across the four data sets, while the hit rate of the individual-level model varies between 67.9 and 70.4 percent. Actually, the predictions of the three-type model are even superior to those of the individual-level model in two of the four data sets. Moreover, the three-type model also predicts substantially better than the representative agent model, beating the predictive accuracy of the latter in each of the four data sets. Thus, similar to the insights from the precision-parsimony frontier, we find that the three-type model leads to a substantial improvement over the representative agent model, while exceeding three types yields

---

[4]The utility function allows for all motives discussed in the literature on distributional preferences such as selfishness, envy, concern for poorer individuals, aversion against disadvantageous inequality, concern for the total payoff, etc.

[5]Note that the predictions of the representative agent model are based on the estimation of two distributional preference parameters, while the three-type model uses six estimated parameters (two for each type). This contrasts sharply with the individual-level model, which uses a set of estimated parameters (2N parameters) that is orders of magnitude larger (e. g., 934 estimated preference parameters for the 467 individuals in the Panel-2017 data set).

[6]The hit rate is the percentage of predicted choices for the prediction set that coincides exactly with individuals' actual choices. Our results are qualitatively similar if we use the mean squared errors of subjects' actual choices from the predicted choices, but the hit rate has the advantage of being simpler and easier to interpret.

only very few additional benefits in terms of predictive accuracy.[7]

But how good is the predictive ability of our three-type model compared to a state-of-the-art machine learning tool? To answer this question, we used regularized gradient boosting trees (rGBT)—a machine learning technique that is widely used in computer science and has been demonstrated to outperform alternative machine learning models in various prediction scenarios because of its iterative error correction mechanism (Shwartz-Ziv and Armon, 2022). Thus, we train rGBT on subjects' decisions on the estimation set and predict their decisions both for the estimation set (within-sample predictions) and for the choice situations in the prediction set (out-of-sample predictions).[8]

A key difference between the structurally estimated three-type approach and rGBT is that the structural model relies on an explicit theory of how individuals make trade-offs between their own and others' payoffs. It uses the estimated values of the types' trade-offs (incorporated in the estimated parameters of the types' utility functions) to make predictions, while rGBT is a "black-box" optimized for predicting outcomes in situations similar to those of the training set. However, the nature of out-of-sample predictions is that the situations used in the predictions may differ from those used to estimate or train the model. Thus, differences in the out-of-sample predictive accuracy between the two approaches also inform us about the potential value of economic theory for prediction purposes: If rGBT outperforms the three-type model in out-of-sample predictions, then the economic theory underlying distributional preference models may be of limited value for predictive purposes. If, in contrast, the three-type model has a better out-of-sample predictive ability, then the structural model not only enables insights into how subjects make trade-offs between their own and others' payoffs in *known* decision situations, but it also provides a superior understanding of how they make these trade-offs in *new, yet unknown*, situations—suggesting that the model is more transferable to new situations than the

---

[7]Note that relaxing parsimony by allowing for four or five types also does not increase predictive power because even if we increase to four or five clusters, almost all individuals remain in the three clusters identified in the three-type model. Thus, the hit rates vary negligibly relative to the three-cluster approach.

[8]Regularization applies penalties to more complex models to prevent overfitting. Gradient boosting trees has previously also been successfully employed in economics (see, e.g., Chalfin et al., 2016; Einav et al., 2018; Deryugina et al., 2019).

rGBT.

We find that the rGBT makes superio*r within-sample predictions, with hit rates varying between 96.8 and 99.9 percent*. This was expected, since rGBTs are optimized to excel at this task. However, the *three-type model far outperforms rGBT when it comes to out-of-sample predictions*. As mentioned above, the three-type model's out-of-sample hit rate varies between 66.5 and 72.5 percent, while the out-of-sample hit rates of rGBT varies between 27.9 and 33.0 percent only. This finding—together with the results mentioned on the previous pages—lends strong support to an economic approach towards distributional preferences that is based on a parsimonious characterization of behavioral heterogeneity at the type-level.

Our paper makes several contributions. First, while economists often allude informally to the desirability of parsimonious models, we are not aware of contributions that explicitly address the trade-off between parsimony and descriptive accuracy. Our application of the DP-means approach makes it possible to empirically quantify this trade-off in terms of the precision-parsimony frontier, allowing us to make decisions about the best level of parsimony in a principled and empirically informed way. We apply our method to identify a parsimonious characterization of behavioral heterogeneity in the domain of distributional preferences but the method is, in principle, also applicable to the domain of risk and time preferences.

Second, we document large precision benefits from only a small reduction in parsimony in all of our four data sets, suggesting that one has to move only slightly beyond the representative agent model to accurately account for the bulk of the prevailing heterogeneity in the population. This view is reinforced by the existence of a kink in the precision-parsimony frontier at three behavioral clusters which implies rather small accuracy benefits beyond three clusters and rather large accuracy losses below three clusters.

Third, our approach enables us to demonstrate that three fundamentally different behavioral types capture the key properties of the population's distributional preferences and to characterize these properties in terms of the prevailing distributional preference models: (i) inequality aversion, (ii) altruism, and (iii) predominant selfishness.

Fourth, we demonstrate that our parsimonious three-type model generates sub-

stantially better predictions than a representative agent model and displays the same predictive accuracy as *n*-types models based on the estimation of each individual's utility function. Thus, instead of knowing an individual's precise preference parameters, it is only necessary to know the individual's type assignment and the type's preference parameters to make quantitative predictions.

The third and the fourth points relate our paper to the literature on preference estimation with finite mixture models (e.g., Bardsley and Moffatt, 2007; Bruhin et al., 2010; Iriberri and Rey-Biel, 2011, 2013; Conte and Moffatt, 2012; Conte and Levati, 2014; Breitmoser, 2013; Bruhin et al., 2018; Burghart et al., 2020) and to the literature on the properties and the stability of social preferences (Andreoni and Miller, 2002; Bellemare et al., 2008, 2011; Fisman et al., 2007, 2015, 2023; Chuang and Schechter, 2015; Carlsson et al., 2014) These papers differ in many dimensions from ours, but the most important difference is perhaps the fact that they do not examine the trade-off between parsimony and precision, which is one of the core questions of our paper. Instead, they either estimate each individual's utility function or estimate finite mixture models that rely on ex-ante assumptions on the structure of behaviors or preferences. This contrasts with our nonparametric approach where the behavioral types emerge endogenously and without any assumptions on utility functions or pre-existing types.

Finally, our paper is also related to a small emerging literature that (i) compares the (out-of-sample) predictive ability of economic models to that of machine learning algorithms (e.g. Camerer et al., 2019; Fudenberg and Puri, 2021; Plonsky et al., 2019, 2017) and (ii) investigates the extent to which economic models are more complete (Fudenberg et al., 2022) and transferable to other domains (Andrews et al., 2022).

The remainder of our paper is structured as follows. Section 2 describes our experimental design and Section 3 studies the fundamental properties of distributional preferences. This section examines whether we can already find first hints for the existence of behavioral types at the purely descriptive level. Then we apply the DP-means algorithm to each of our four data sets to uncover the fundamental aspects of population heterogeneity in a more principled and rigorous way. Section 4 examines the relative predictive power of a type-based empirical approach by comparing it to the representative agent model, an individual-level model, and a state-of-the-art machine learning model. Section 5 concludes the paper.

2   Experimental design

2.1   Measuring social preferences

We measured subjects' social preferences using a series of incentivized money allocation tasks in which participants had to decide how to allocate experimental currency units (ECUs) between themselves and another anonymous participant in the study.[9]

The literature on the measurement of social preferences has predominantly relied on choice situations where the decision maker can *increase* the recipient's payoff by giving up some of her own payoff to identify other-regarding behavior (see, e.g., Fisman et al., 2017). In our study, we use a combination of choice situations in which the decision maker can pay to *increase* the other participant's payoff (negatively sloped budget lines in "self payoff–other payoff" space), *and* choice situations in which she can pay to *decrease* the other participant's payoff (positively sloped budget lines in the "self payoff–other payoff" space). We also use positively sloped budget lines because negatively sloped budget lines alone do not allow us to identify a wide range of potentially relevant other-regarding behaviors. For example, inequality aversion Fehr and Schmidt (1999) implies that individuals are not only willing to sacrifice some of their own payoff to increase the payoff of those worse off (aversion to advantageous inequality), but that they are also willing to sacrifice resources to *decrease* the payoff of those who are better off (aversion to disadvantageous inequality). While negatively sloped budget lines allow us to identify the former motive, the latter requires positively sloped budget lines. Similarly, envious or spiteful individuals can only be identified when they have the chance to reduce others' payoff at a cost to themselves. We address these identification issues by eliciting subjects' distributional choices on a set of both negatively sloped and positively sloped budget lines.

Figure 1 depicts the 12 budget lines that we use to identify subjects' distributional preferences. These choice situations systematically vary the cost of redistribution and its impact on joint payoffs, thereby allowing us to identify a wide variety of other-

---

[9]To avoid issues related with reciprocity, we made it clear to our participants that they would in no way be affected by any decision the other participant makes. This was explained as follows in the instructions: *"The other participant will take part in another part of the study. Anonymity between you and the other participant is guaranteed, i.e., neither you nor the other person will ever learn about each other's identity. Moreover, the other participant will not take decisions that affect you, i.e., you will not be affected in any way by the other participant's decisions."*

regarding behaviors. In addition to these choice situations, we also presented our subjects with a set of additional budget lines that we use for out-of-sample predictions (more details on them is provided in Section 4).

Figure 2a illustrates how a typical choice situation was presented to participants. We represented the available choices numerically and graphically in order to make the trade-offs and the associated payoff implications transparent. For each choice situation, there were always seven interpersonal allocations (labeled by 1 to 7) – all located on a budget line. Each available allocation consisted of a specific distribution of ECUs between the participant (bars labeled by "You receive") and the other person (bars labeled by "Other person receives"). Figure 2b plots the budget line corresponding to the example depicted in Figure 3a in the "self-payoff (w^s) – other's payoff (w^o)" space. In this example, the slope of the budget line is -2, indicating that for every ECU the decision maker gives up, the other participant receives 2 ECUs. Perfect equality in payoffs can be achieved by choosing allocation 4.

Figure 1: Measuring distributional preferences with a money allocation task
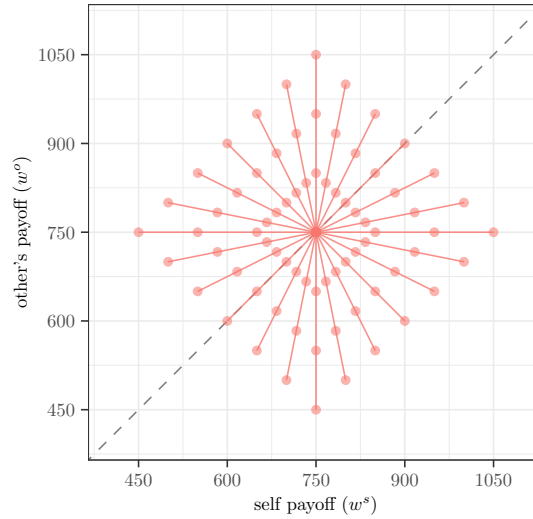
2.2 Samples and implementation

We collected our data in 2017 and in 2020, covering a total of 1731 individuals who are broadly representative of the French and the German language population of Switzerland with respect to age, gender, and region. 348 respondents were surveyed

Figure 2: Example of a choice situation

(a) Decision screen

(b) Budget line
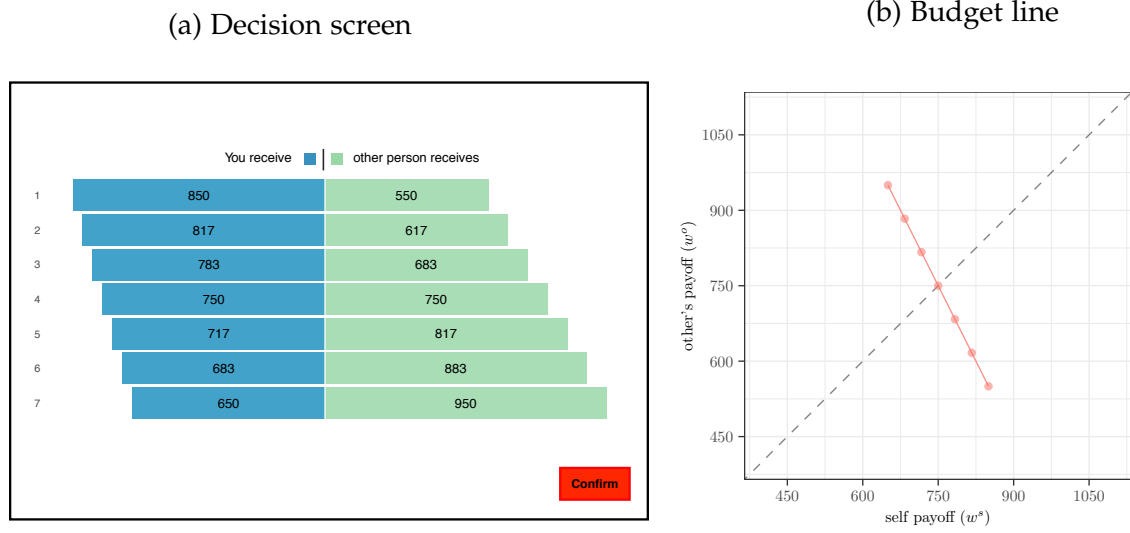
in 2017 only (*Only-2017 sample*), 916 respondents were surveyed in 2020 only (*Only-2020 sample*) and 467 respondents were surveyed *both* in 2017 and in 2020 (*Panel sample*), yielding four separate datasets where we can investigate the distribution of social preferences.[10] We display the main descriptive statistics of each data set in Table A.4 in the Appendix.

The procedure and the implementation of the experiment was similar for all the subjects and across data sets. The experiment was computerized and conducted on-line in collaboration with the LINK Institute, the leading company for high-quality market research in Switzerland. In all samples, participants were paid a show-up fee for taking part in the study until the end. In addition, we also incentivized participants' choices in the money allocation task by paying out one of their (randomly drawn) decisions. The exchange rate between points in the money allocation task and Swiss Francs was always 40 points per CHF.

---

[10]There are, in total, four datasets in which subjects could make independent allocation decisions because the panel sample consists of two subsamples: one comprising panel subjects' decisions in 2017 (*Panel-2017*), and another one comprising panel subjects' decisions in 2020 (*Panel-2020 sample*).

3 The fundamental properties of distributional preferences

In this section, we explore the behavioral properties of individuals' distributional preferences. Previous evidence (see, e.g., Andreoni and Miller, 2002; Bellemare et al., 2008; Fisman et al., 2007, 2017; Kerschbamer and Müller, 2020; Bruhin et al., 2018; Cappelen et al., 2007) suggests that there is considerable heterogeneity in distributional preferences. We are primarily interested in the question whether the distribution of preferences can be captured parsimoniously with a small number of *behavioral types* that exhibit fundamentally distinct social preferences. To that end, we examine whether the population consists of distinct clusters of individuals.

We approach this task in two steps. We start with a descriptive analysis of subjects' behavior in the money allocation task. This analysis could, in principle, already provide first hints on the existence of clusters of individuals with clearly distinct behavioral properties. We then proceed with a more rigorous approach in which we apply a Bayesian nonparametric clustering algorithm—the Dirichlet Process (DP) Means algorithm (Kulis and Jordan, 2012)—to examine behavioral heterogeneity in our data sets.

The DP-means algorithm has several appealing properties that fit our purposes. *First*, because the algorithm can be applied to individuals' "raw" allocation data, it does not require any assumptions about the *behavioral properties* (e.g., utility functions) of the population under study. Instead, the algorithm is completely agnostic regarding the concrete behavioral regularities that emerge, i.e., it is entirely driven by the characteristics of the data. Thus, our approach differs from previous work (e.g., Bellemare et al., 2008; Fisman et al., 2015, 2017; Bruhin et al., 2018) that characterized preference heterogeneity on the basis of ex-ante assumptions on feasible utility functions. *Second*, the algorithm allows the identification of behavioral heterogeneity without ex-ante assumptions on the *number* of existing clusters. Instead, it makes the trade-off between parsimony and descriptive accuracy explicit by requiring the researcher to determine a level of precision with which individuals are assigned to clusters. Once the level of precision is fixed, the algorithm endogenously determines the number of clusters and assigns all individuals to one of the emerging clusters

(with probability one[11]) in a way that minimizes the sum of individuals' deviations (measured in terms of the squared Euclidean distance) from their clusters' centroids.

A *third* attractive property of the DP-means algorithm is that it nests the full range of types, from a representative agent setting (i.e., a single data-generating process) to a model where all individuals differ in their preferences (i.e., n data-generating processes with as many types as there are individuals). *Fourth*, because the algorithm can be directly applied to individuals' observed behavior and does not rely on ex-ante assumptions about behavior, it does not require any assumptions on the *structure of utility noise* or the *structure of behavioral errors* (i.e., the error terms).[12]

3.1   Identifying behavioral heterogeneity at the descriptive level

The key properties of distributional preferences pertain to individuals' willingness to pay to *increase* others' payoff *and* their willingness to pay to *decrease* others' pay-off. Therefore, when examining the potential existence of behavioral types at the descriptive level, we search for clusters of individuals who display distinct but typical patterns in their willingness to pay to increase and decrease others' payoffs in different distributional situations.

For this purpose, we plot each individual's *modal* choice across the negatively sloped and across the positively sloped budget lines. An individual's modal choice on negatively sloped budget lines informs us about their willingness to pay to increase the other's payoff, whereas their modal choice on positively sloped budget lines informs us about her willingness to pay to decrease the other's payoff.[13]

We depict subjects' modal choices on positively sloped and on negatively sloped budget lines separately for each of our four data sets in Figure 3. For each budget line,

---

[11]This distinguishes DP-means from alternative approaches that probabilistically classify individuals into clusters (such as the finite mixture models used, e.g., in Bruhin et al., 2010, Burghart et al. (2020) or Bruhin et al., 2018).

[12]While the advantages of being able to infer behavioral clusters and individuals' assignment to clusters without structural assumptions on behavior are relatively transparent, the advantage of avoiding assumptions about the structure of the error term (or utility noise) may seem less obvious. However, it has been shown in the domain of risk preferences that assumptions about the utility noise in random utility models are not innocuous. For instance, Buschena and Zilberman (2000) showed that the assumptions on the error term are decisive for whether expected utility theory or non-expected utility models best capture the data.

[13]We focus on the mode because it is less susceptible to random influences and outliers. Note that we get very similar results if we instead use subjects' median choices.

we label the own-payoff maximizing allocation by $z = 1$, the own-payoff-minimizing allocation by $z = 0$, and the payoff-equalizing allocation by $z = 0.5$. The other four available allocations on each budget line are equidistantly placed between 0–0.5 and 0.5–1, respectively.

Strikingly, *the same three distinct behavioral agglomerations emerge in all four data sets*:

i) One group of individuals is located at $z = 0.5$ for both positively and negatively sloped budget lines. These individuals seem to be primarily motivated by equality as they tend to choose the equal-payoff allocation, irrespective of whether this means paying to reduce the other person's payoff (positively sloped budget lines) or paying to increase the other's payoff (negatively sloped budget lines).

ii) Another group of individuals is located at $z = 1$ for positively sloped budget lines and $z = 0.5$ for negatively sloped budget lines. These subjects tend to equalize payoffs on negatively sloped budget lines but to maximize their own payoff on positively sloped budget lines. Overall, they do *not* seem to be willing to pay to *reduce* the other subject's payoff *for the sake of equality*, but they behave in an altruistic way when on negatively sloped budget lines, i.e., they are willing to give up some of their own payoff to increase the other's payoff.

iii) The last group of individuals is located at $z = 1$ for both negatively and positively sloped budget lines. These subjects tend to predominantly maximize their own payoff irrespective of the consequences of this choice for the other participant, i.e., they appear to make primarily self-interested choices.

Overall, Figure 3 provides a consistent message. First, it suggests the existence of clusters of individuals with a clear behavioral interpretation – one cluster of individuals that primarily care for equality, another cluster that appears motivated by altruistic concerns but never reduces others' income, and a predominantly selfish cluster. Second, the observed clustering supports the view that behavioral heterogeneity may indeed be well captured by a *parsimonious* number of types. Third, the figure indicates that the type distribution is quite stable, i.e., that the *same* qualitative behavioral types emerge *across all data sets* (Figures 3a to 3d) and *over time* (Figures 3a and 3b). In the next section, we characterize behavioral heterogeneity more formally

using Dirichlet Process Means. This enables us to examine the conclusions suggested in Figure 3 more rigorously.

Figure 3: Subjects' modal choices on negatively sloped and on positively sloped budget lines

(a)

(b)

(c)

(d)
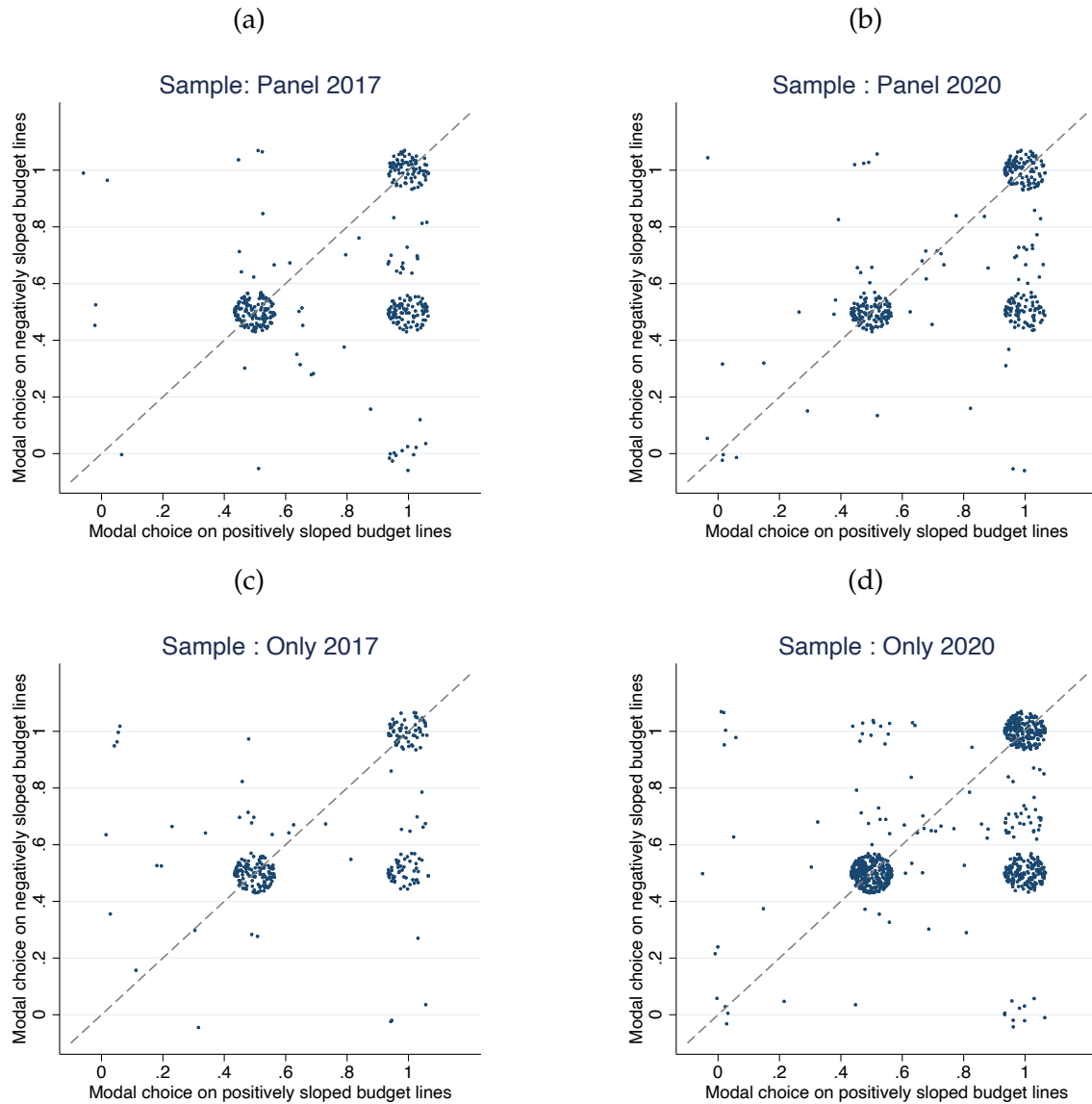
Note: In all figures, we depict subjects' modal choices among negatively sloped budget lines and among positively sloped budget lines. Each dot represents one individual. Dots are jittered in order to make identical modal choices of individuals visible. For each budget line, $z = 1$ indicates an own-payoff maximizing choice, $z = 0$ indicates an own-payoff minimizing choice, and $z = 0.5$ indicates a payoff-equalizing choice. Panel (a) is constructed using panel subjects and their 2017-choices. Panel (b) is constructed using panel subjects and their 2017-choices. Panel (c) is constructed using the choices of individuals who only participateed in the 2017 study. Panel (c) is constructed using the choices of individuals who only participateed in the 2020 study. Note that if we replace individuals' modal choices by their median choices, very similar behavioral agglomerations emerge.

3.2 Identifying behavioral heterogeneity with the DP-means algorithm

The descriptive analysis in the previous section strongly suggests that our data may be populated by three stable behavioral types. However, this analysis only takes subjects' modal choices into account, whereas a more rigorous analysis should be based on all the choices subjects make. In addition, the descriptive analysis ignores the precision costs of parsimony that result when restricting attention to a small number of clusters. Moreover, the descriptive analysis only suggests that there are three stable behavioral agglomerations, but it does not assign each individual to one of the behavioral types. Since individuals' choices are often not perfectly aligned with one of the type's typical choices, one needs an explicit metric that enables the assignment of individuals to types. Finally, the descriptive analysis also fails to provide a principled way to assess whether a higher, yet still parsimonious, number of behavioral types (e.g., 4 or 5 types) exists. We address these issues by applying the Dirichlet Process Means algorithm developed by Kulis and Jordan (2012).

In the following, we first describe the details of this algorithm and outline its specificities and properties. We also discuss its advantages over the better known *k*-means algorithm. We then apply this algorithm to our data sets.

3.2.1 The Dirichlet Process (DP) means algorithm

To apply the DP-means algorithm, we consider all the choices subjects make in the 12 budget lines depicted in Figure 2. An individual is thus represented by its *allocation profile*, i.e., the set of all 12 budget allocations normalized to the unit interval, where 1 refers to the own-payoff-*maximizing* choice and 0 refers to the own-payoff-*minimizing* choice on the budget line.[14] We search for clusters of individuals in the 12-dimensional unit space spanned by the 12 budget lines, with each individual being represented as a single data point (her allocation profile) in that space. The output of the algorithm is the number of clusters that emerge in the allocation space, and the individuals' assignment to these clusters. We call a cluster of individuals a *be-*

---

[14]As a convention, and without loss of generality, we set the normalized allocation to 1 if the other's payoff is maximized for the (vertical) budget line with infinite slope.

havioral (or preference) type if the behavior of the individuals in that cluster has a clear interpretation in terms of subjects' distributional preferences.

The algorithm characterizes clusters by a type-specific mean allocation vector (the centroid) to which individuals lie close (in terms of the squared Euclidean distances between individual allocation profiles and the nearest centroid). The DP-means algorithm requires specification of a parameter λ that has two intuitive meanings. First, λ represents the maximal allowable (squared Euclidean) distance between individuals belonging to a cluster and that cluster's centroid. Thus, if an individual's distance from a cluster's centroid is larger than λ, the individual does not belong to that cluster. This means that a lower value of λ will tend to increase the accuracy with which individuals are assigned to clusters (since they are constrained to lie closer to their closest centroid), but it will also increase the number of clusters (thereby reducing parsimony). λ thus makes the tradeoff between descriptive accuracy and parsimony explicit. Second, as shown by Kulis and Jordan (2012), λ also represents the proportional cost of adding an additional cluster (the 'cost of complexity') in the objective function (1) described below.

The algorithm works as follows: We initially start with a single centroid specified as the global mean vector of all allocation profiles, i.e., the mean of all observations in the budget allocation space. At this stage, all individuals (i.e., all data points) are assigned to a single representative agent. We then refine by iterating over the following two steps:

1. We go over the list of individual allocation profiles in the allocation space and check whether any of the squared Euclidean distances to the centroids exceeds the maximally allowable distance λ for each individual. If this is the case, we open a new cluster with the allocation profile that exceeds λ as its centroid. Otherwise, we assign the data point to its nearest cluster.

2. We collect the subjects assigned to the same clusters and update the centroids by computing the mean vector of all observations belonging to a cluster.

These two steps are repeated until convergence is reached, i.e., until there is no more change in subjects' assignments. As Kulis and Jordan (2012) demonstrate, this

18

iterative procedure monotonically decreases the following objective at each iteration:

$$\min_{\{g_c\}_{c=1}^{k}} \sum_{c=1}^{k} \sum_{x \in g_c} \|x - \mu_c\|^2 + \lambda k, \tag{1}$$

where x denotes an individual's allocation profile, μ_c denotes the centroid of cluster c, g represents the cluster partitioning of individuals, i.e., an assignment of each individual's allocation profile x to one of the clusters g_c, and k denotes the number of clusters.

The objective function described in Equation (1) is equivalent to the k-means objective, with the exception of the additional penalty term λk. An important advantage of DP-means over *k*-means is that it is better suited to discover the true number of clusters in the data and that it yields clusters of higher quality in the sense that individual observations are more likely to be assigned to the correct cluster (for a more extensive discussion, see Kulis and Jordan (2012) and Comiter et al. (2016)).

As a demonstration of DP-means' capability to recover the true behavioral heterogeneity, we show in Appendix C that three different simulated types with constant elasticity of substitution (CES) utility are detected by the algorithm, even in the presence of utility noise or random choice errors.

3.2.2 Applying the DP-means algorithm to our data sets

To identify a reasonable starting point for the application of the DP-means algorithm, we examine the properties of the precision-parsimony frontier in each of our four data sets. To that end, we calculate the loss in precision—in terms of the mean over all individuals' squared Euclidean deviations from the centroids (MSD)—generated by a more parsimonious clustering. Figure 4 depicts MSD as a function of the number of clusters separately for each of the four data sets.[15]

The figure shows that increasing the number of clusters from one to two, and from two to three is associated with large reductions in individuals' mean deviations from their centroids. In contrast, further increasing the number of clusters to four or

---

[15]To guide our intuition regarding the units of MSD displayed on the vertical axes, it is useful to know that if an individual's allocation profile is on average one allocation away from the centroid, $MSD = 0.333$.

more yields only very small gains in precision. In fact, the gains in precision from increasing the number of clusters below $k = 3$ is almost an order of magnitude higher than the precision gains that accrue form exceeding $k = 3$.[16] This means that starting to examine the behavioral implications of our clustering results at λ-levels that yield three clusters is not only suggested by the descriptive analysis in the previous section; but also by the parsimony-precision trade-off.

When examining the behavioral implications of a cluster structure, we essentially ask two questions. First, we ask whether the clusters allow for a clear and unambiguous behavioral interpretation, i.e., *whether each cluster indeed represents a preference type.* For example, a cluster that would mix up selfish and altruistic individuals would be unsatisfactory as it would be difficult to assign a clean preference interpretation to this cluster. In this context, it is important to recall that the DP-means algorithm is agnostic with respect to the behavioral interpretation of the different clusters; the algorithm only groups individuals according to their observed behavioral similarities, regardless of what these similarities may be. It is the task of the researcher to inspect and interpret the behavior of the individuals in the different clusters identified by the algorithm.

Second, we also ask whether *the same* preference types emerge across the four different data sets, i.e., whether the clusters are *stable across subject pools and over time.* Suppose, for example, that we identified three clusters of (i) envious, (ii) altruistic, and (iii) selfish individuals in one data set, but found that an inequality averse type replaces the altruistic type in another data set drawn from the same population. Then our findings would be less useful for theoreticians who strive for a parsimonious, yet empirically based, modelling of social preferences because no stable type structure would exist in that population. Thus, in addition to choosing an ideal point on the parsimony-precision trade-off, we also care about suitability of the behavioral interpretation and type stability. From an economic viewpoint, these are clearly desirable criteria for describing the preference/behavioral heterogeneity in a population.

---

[16]More precisely, the reduction in MSD when moving from 1 to 2 clusters is between 0.17 and 0.29 units across data sets, and when moving from 2 to 3 clusters it is between 0.12 and 0.24 units across data sets. This contrasts sharply with the reduction in MSD that results from moving beyond three clusters. Moving from 3 to 4 clusters reduces MSD between 0.01 and 0.03 units; and moving from 4 to 5 clusters reduces MSD between 0.01 and 0.04 units. Thus, the precision gains of increasing k when $k < 3$ are discontinuously larger compared to the gains from going beyond $k = 3$.

Figure 4: Mean over all squared deviations of individuals from their centroids (MSD) as a function of the number of clusters.

(a) Sample: Panel-2017

(b) Sample: Panel-2020

(c) Sample: Only-2017

(d) Sample: Only-2020
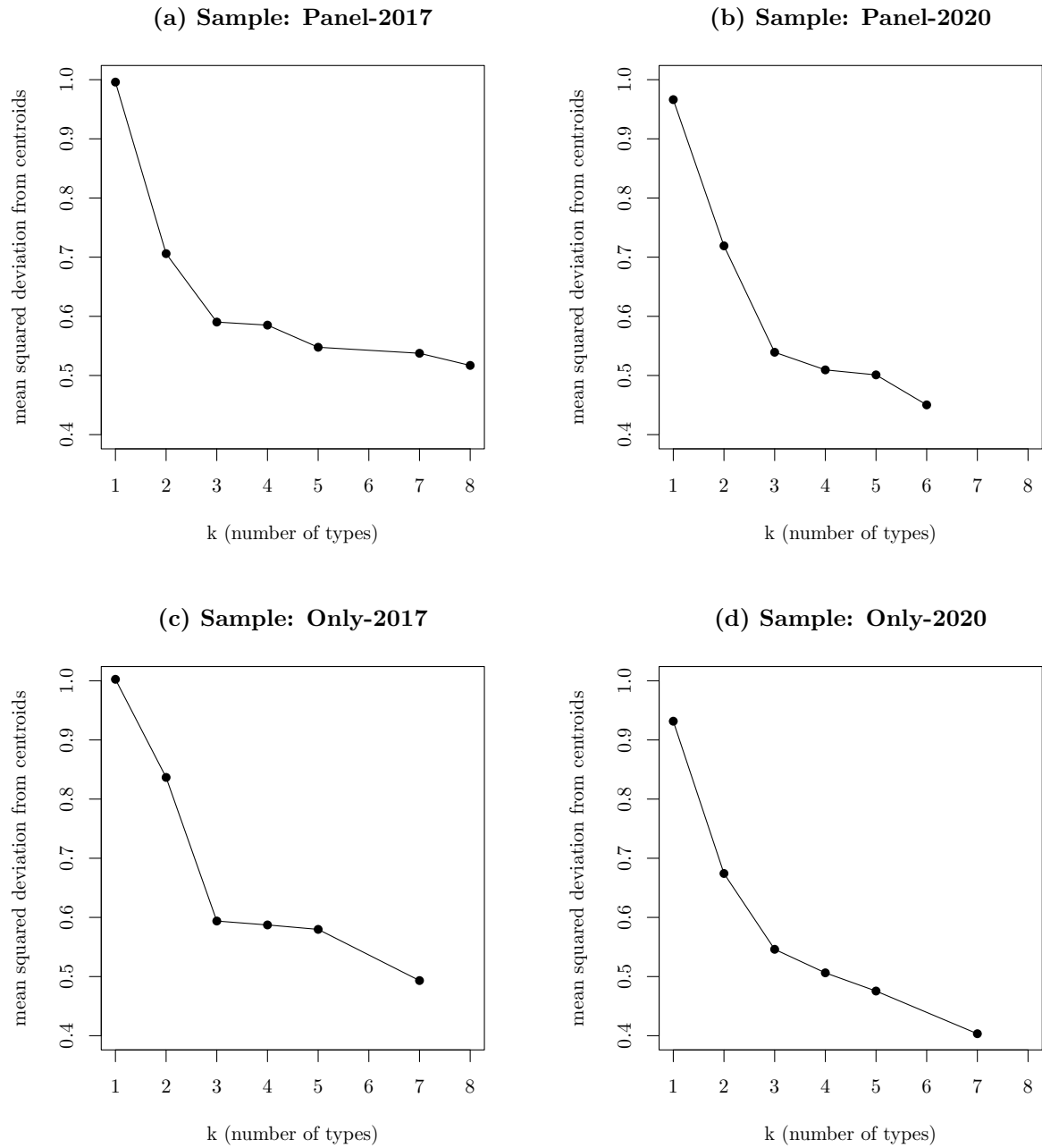
Note: The figure shows the mean over all squared deviation (measured in terms of Euclidean distance) of individuals from their centroids as a function of the emerging number of clusters.

What do we find in each of our four data sets? Table 1 summarizes the outcomes of the DP-means algorithm for this case. The largest cluster (Cluster 1) comprises between 45% and 53% of the subjects. The second-largest cluster (Cluster 2) comprises between 30% and 40% of subjects, and the remaining subjects – always the minority

– are assigned to Cluster 3. For example, roughly half of the subjects (48.18%) in the Panel-2017 data set are assigned to Cluster 1, 41.97% to Cluster 2, and the remainder (8.85%) are assigned to Cluster 3.

Table 1: Distribution of behavioral types across data sets with three clusters

	Panel-2017	Panel-2020	Only-2017	Only-2020
Cluster 1 (Inequality Averse)	48.18%	45.18%	53.74%	45.52%
Cluster 2 (Altruistic)	41.97%	38.76%	29.60%	30.46%
Cluster 3 (Selfish)	9.85%	16.06%	16.67%	24.02%

Note: The table displays the distribution of individuals to three clusters (in percent) that emerges in each of our four data sets. The behavioral interpretation of the clusters (indicated in parenthesis in the leftmost column) is based on the interpretation of each cluster's typical behavior provided in Figure 5 below.

Do these clusters have a clear behavioral interpretation, and to what extent do they correspond to fundamentally distinct behavioral types? Figure 5 provides the answer to these questions: It depicts the relative share of own-payoff minimizing, payoff-equalizing, and own-payoff-maximizing choices, respectively, among negatively sloped and positively sloped budget lines, separately for each identified cluster and for each data set.[17]

Individuals assigned to the first cluster (Cluster 1) predominantly make payoff-equalizing choices. They exhibit a willingness to pay for reducing inequality when this involves increasing the other individual's payoff (budget lines with negative slope) and when it involves decreasing the other individual's payoff (budget lines with positive slope). For this reason, we label this behavioral type as *"inequality averse."* In contrast, individuals in the second cluster display a substantial willingness to pay when the sacrifice involves an increase in the other individual's payoff (i.e., on budget lines with negative slope) but not when it involves a decrease in the other's payoff (i e., on budget lines with positive slope). We therefore label individuals in this cluster as *"altruists."* Finally, individuals in the third cluster make predominantly own-payoff maximizing choices both on budget line with negative and positive slopes. We therefore label them as *"predominantly selfish"*.

---

[17]Recall that subjects could choose among seven different allocations. A choice is classified as own-payoff minimizing (own-payoff maximizing) if it belongs to the two choices that give the subject the lowest (highest) payoff. It is classified as payoff-equalizing if it implements perfect equality or one of its nearest neighbouring allocations.

Figure 5: The distribution of choices for positively and negatively sloped budget lines in each cluster and each data set.

Relative share of
- own-payoff minimizing choices
- payoff-equalizing choices
- own-payoff maximizing choices
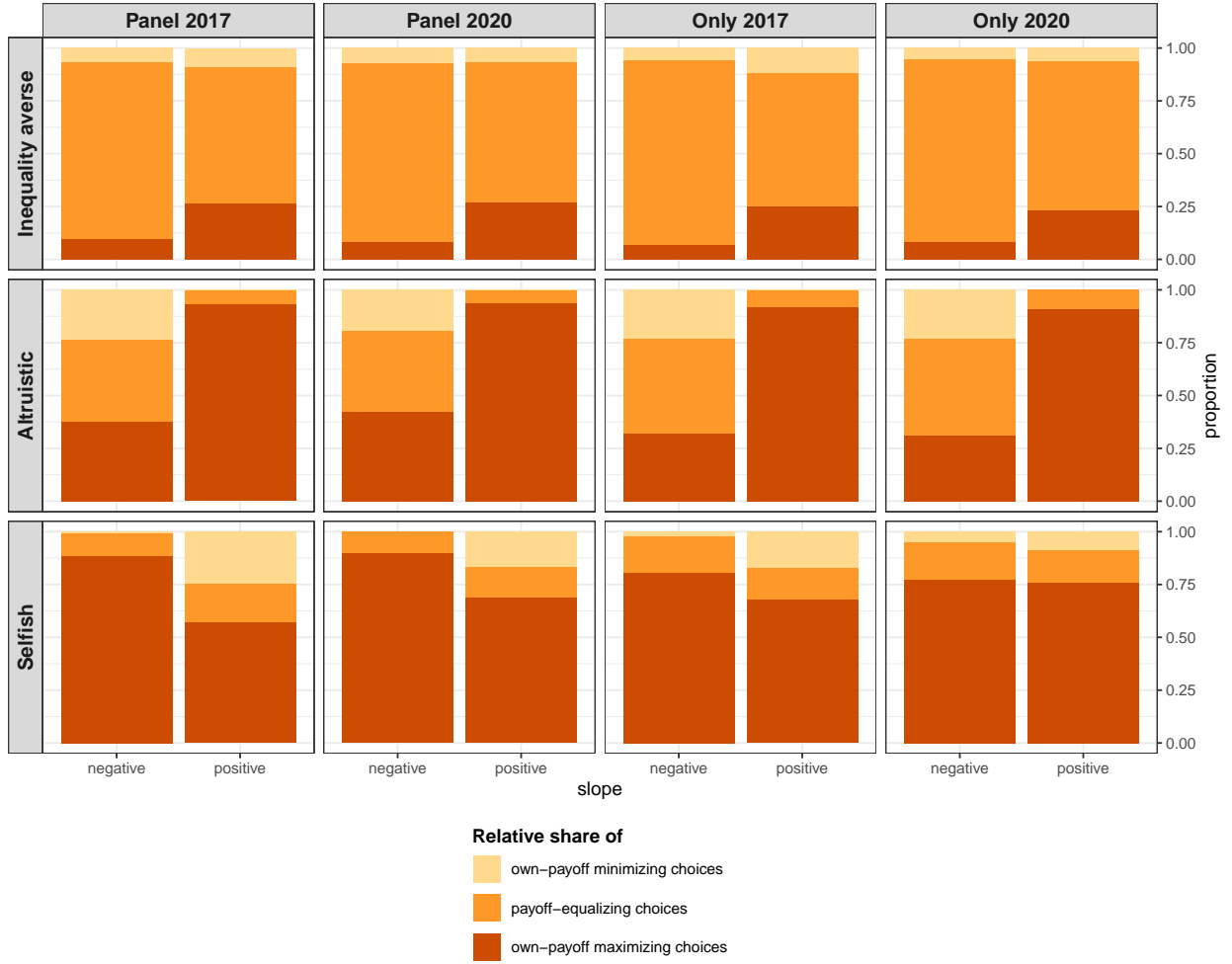
Three aspects of these findings are particularly remarkable. First, the behavioral interpretation of each of the three clusters is similar across the four data sets: subjects are either characterized as being predominantly inequality averse, predominantly altruistic, or predominantly selfish. Second, the aggregate distribution of behavioral types is rather stable across the samples. In all the data sets, the largest cluster comprises inequality averse individuals, a large yet smaller share of individuals is assigned to the altruistic type, and a minority of subjects are assigned to the predominantly selfish type. Third, the aggregate distribution of types is also remarkably stable over time. This can be seen by comparing the distributions of types for the panel sample in 2017 and in 2020.

These findings are particularly remarkable when considering that the DP-means

algorithm i) is agnostic with respect to the behavioral interpretation of the types that populate each sample, and ii) does not put any constraints on the distribution of behavioral types.

3.3 Are there really three behavioral types?

What happens if we allow for a larger number of clusters by reducing the maximal deviation allowed between an individual and their centroid? Will this lead to the emergence of new *empirically important* and *fundamentally distinct* behavioral types that remained hidden under the three-cluster characterization? Table 2 displays the distribution of types in the four data sets when λ is lowered enough so that a four-type distribution emerges. This analysis reveals that *no new meaningful behavioral types appear*.

In three out of four data sets (columns 1-3), the fourth cluster consists of less than 2% of the individuals and is thus quantitatively negligible.[18] Moreover, the few individuals that populate the fourth cluster display rather random and hard to interpret behaviors. In contrast, the remaining three clusters show behavioral regularities that are very similar to those documented with the three-type distribution: a majority of inequality averse subjects, a large group of altruistic individuals, and a minority of predominantly selfish individuals. In addition, we show in Appendix B.1 that virtually all the individuals from these data sets remain assigned to the same behavioral type regardless of whether $k = 3$ or $k = 4$. This means, for example, that an individual assigned to the inequality averse type when there are three clusters almost always remains assigned to the same type when there are four clusters. Thus, individuals' assignment to behavioral types remains robust to an increase in the number of clusters allowed to emerge in these data sets.

Our findings for the remaining data set (Only-2020) appear slightly different at first sight, but not when we examine them more carefully. In this sample, we find that about one-third of the subjects is assigned to the altruistic type (Cluster 2), and about 22% are assigned to the selfish type (Cluster 3). These shares are remarkably

---

[18]The fourth cluster is populated by only one individual in the Panel-2017 sample and in the Only-2017 sample. Even in the Panel-2020 sample, it is populated by only 9 individuals and thus cannot be considered an empirically relevant and fundamentally distinct behavioral type.

Table 2: The distribution of behavioral types with four clusters

	Panel-2017	Panel-2020	Only-2017	Only-2020
Cluster 1	IA (47.97%)	IA (44.11%)	IA (53.74%)	IA-1 (28.28%)
Cluster 2	Altruistic (41.97%)	Altruistic (38.76%)	Altruistic (29.31%)	Altruistic (30.02%)
Cluster 3	Selfish (9.85%)	Selfish (15.20%)	Selfish (16.67%)	Selfish (22.16%)
Cluster 4	– (0.21%)	– (1.93%)	– (0.29%)	IA-2 (19.54%)

Note: The table displays the distribution of individuals to four clusters (in percent) in each of our four data sets. IA indicates inequality averse behavior, and IA-1 (IA-2) indicate the two clusters of inequality averse subjects. "Selfish" indicates the primarily self-interested behavioral type.

similar to the proportions documented under the three-types specification in Table 1, where 30.46% of the Only-2020 sample is assigned to the altruistic type and 24% is assigned to the predominantly selfish type. The remaining subjects are the inequality averse, who are divided into two separate clusters (Cluster 1 and Cluster 4). This interpretation is supported by Table 3 below, which displays the transition of individuals between types for this data set. The table confirms that the inequality averse type under k =3 is divided up into two inequality averse sub-types under $k = 4$, while other instances of type transitions are extremely rare. For example, only 5 individuals (out of 279) belonging to the altruistic cluster under k = 3 switch to one of the other types when $k = 4$. It is further supported by Figure A.1 in Appendix B.1 which depicts subjects' choices among negatively sloped and positively budget lines. Thus, taken together, the evidence suggests that no new meaningful behavioral types emerge if we allow for four clusters, and that an individual's assignment to types remains very stable when moving from three to four clusters.

Table 3: Transition of individuals between behavioral types in the Only-2020 data set

		k = 4 clusters				
		Inequality averse (1)	Inequality averse (2)	Altruistic	Predominantly selfish	Total (%)
k=3	**Inequality averse**	240	176	1	0	417 (45.52%)
clusters	**Altruistic**	5	0	274	0	279 (30.46%)
	Predominantly selfish	14	3	0	203	220 (24.02%)
	Total (%)	259 (28.28%)	179 (19.54%)	275 (30.02%)	203 (22.16%)	916 (100%)

What happens if we are willing to become even less parsimonious and allow for five clusters? In Appendix B.2, we show that no new empirically relevant and

fundamentally distinct behavioral type emerges when allowing for a finer partition of the data.

While we have seen that increasing precision by reducing parsimony does not bring novel insights compared to a three-type model, one might wonder whether it would be possible to increase parsimony by restricting heterogeneity to two clusters only. Would this more parsimonious description of behavioral heterogeneity still yield a clean and stable distribution of types in the population, or would we lose important insights relative to three clusters?

We display the results of this exercise in Table 4 below. The table shows that when we allow for only two clusters, the algorithm systematically merges fundamentally distinct behavioral types together. As a result, most clusters do not have a clear behavioral interpretation when $k = 2$. We therefore label these clusters as mixtures of incompatible preferences ("MIP"). For example, in the Panel-2017 data set, cluster 1 (under $k = 2$) comprises subjects from the selfish and the inequality averse type (under $k = 3$). We show this explicitly in the first row of the transition matrix in Table 5 below. In the Panel-2020 data set, the selfish type disappears again under $k = 2$ because it is now absorbed by the altruistic type in Cluster 2 (see Table A.13 in Appendix B.3), thereby also preventing any clean interpretation of clusters in terms of behavioral types. Similar difficulties arise in the two remaining data sets, as described in Tables A.14 and A.15 of Appendix B.3.

These results highlight that requiring a higher degree of parsimony so that only two clusters emerge yields an unsatisfactory characterization of the behavioral heterogeneity as i) it systematically makes an important type disappear, ii) it makes a clean behavioral interpretation of the emerging clusters impossible, and iii) it undermines the between-samples stability of the distributions of behavioral types.

Table 4: The distribution of behavioral types with two clusters

	Panel-2017	Panel-2020	Only-2017	Only-2020
Cluster 1	MIP (55.89%)	IA (46.25%)	MIP (82.76%)	MIP (54.04%)
Cluster 2	Altruistic (44.11%)	MIP (53.75%)	Selfish (17.24%)	MIP (45.96%)

Note: The table displays the distribution of individuals to two clusters (in percent) in each of our four data sets. The behavioral interpretation of the clusters is based on the information provided by Tables 5 and Tables A.13 to A.15 in Appendix B.3. "IA" indicates inequality averse behavior. "MIP" indicates a mixture of incompatible preferences.

Table 5: Transition of individuals between types in the Panel-2017 data set

| | | k = 3 types | | | |
		Inequality averse	Altruistic	Predominantly Selfish	Total (%)
k=2	**MIP**	219	0	42	261 (55.89%)
types	**Altruistic**	6	196	4	206 (44.11%)
	Total (%)	225 (45.18%)	196 (41.97%)	46 (9.85%)	467 (100%)

Altogether, these findings strongly support the conclusion that the behavioral heterogeneity in our data sets is best represented by *three* fundamentally distinct, and stable, preference types with a clear behavioral interpretation: a densely populated inequality averse type, a smaller yet still large altruistic type, and a third type comprising the minority of selfish subjects.

4   The predictive power of behavioral types

In the previous sections, we have shown that a three-type characterization of behavioral heterogeneity is strikingly stable both across data sets and over time. We have also shown that allowing for more than three types does not bring important new insights in terms of uncovering fundamental behavioral heterogeneity, and that restricting the number of types to less than three results in a substantial loss of information. But how well does such a parsimonious three-types model predict *individual* behavior in novel choice situations? Does such a high degree of parsimony impair the model's predictive ability, or does it predict as well as a model that allows for more heterogeneity? In other words, what is the cost of parsimony in terms of predictive ability?

4.1   Does parsimony impair out-of-sample predictions?

In this section, we investigate the out-of-sample predictive ability of our type-based characterization of preference heterogeneity. We compare it with the predictive ability

of (i) a representative agent model, and (ii) a model with individual-level heterogeneity. These two alternative specifications represent the most extreme characterizations of behavioral heterogeneity: The representative agent model assumes that all agents can be described by the same set of parameters and thus rules out any form of behavioral heterogeneity, whereas the model with individual-level heterogeneity allows all individuals to differ from each other. In general, one would expect that empirical models that capture differences in individuals' preference parameters also have a higher ability to predict *individuals*' behavior. At the same time, however, stochastic factors also affect individual behaviors. This randomness may have less impact on type-based or representative agent-based predictions than predictions based on individuals' utility functions. This follows from the fact that parts of this randomness may cancel out when individuals are pooled together. This raises two important questions: (i) Is our parsimonious three-type model indeed superior to the representative agent model? and (ii) How much worse is the predictive ability of the three-type model compared to a model allowing for individual-level differences?

To answer these questions, we apply the random utility approach with a two-parameter utility function that nests altruistic, inequality averse, and selfish preferences. Because we already know from our non-parametric analysis that these three types constitute our population's preferences, we can be confident that the application of our parametric distributional preference model is not misspecified. The parameters of the distributional random utility model are estimated using subjects' choices on the twelve budget lines shown in Figure 1. The estimated preference parameters allow us to make *quantitative* predictions for out-of-sample budget lines that were not used for parameter estimation.

Figure 6 below shows the budget lines (choice situations) for which we make out-of-sample predictions. Note that the figure only contains budget lines that do *not* cross the 45-degree line ("non-center budget lines"), while the budget lines used for the identification of types and for estimating the parameters all crossed the 45-degree line symmetrically and were thus centered (Figure 1). This means, for example, that a sufficiently inequality averse subject who chooses the equal payoff allocation in the middle of a negatively sloped *centered* budget line is predicted to *minimize* her payoff on a negatively sloped budget line that is located completely *below* the 45-degree

28

line (i.e., on a budget line in the advantageous payoff domain), since this minimizes inequality. Likewise, such an individual is predicted to *maximize* her payoff on a negatively sloped budget line that is located completely *above* the 45-degree line (i.e., on a budget line in the disadvantageous payoff domain), since this also minimizes inequality. Thus, because we predict behavior for non-centered budget lines on the basis of behaviors (and model estimates) on centered budget lines, the predictions often deviate strongly from the behaviors of subjects on centered budget lines. This means that the out-of-sample predictions constitute a serious predictive challenge for the estimated empirical models.

Figure 6: Choice situations used for out-of-sample predictions
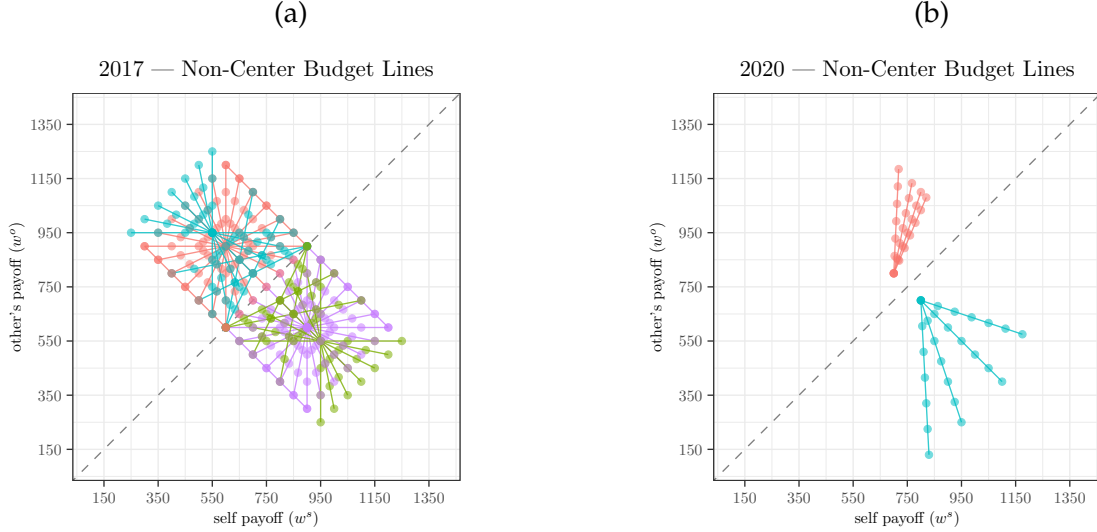
Note: Figure 6a shows the non-center budget lines for which the Only-2017 subjects and the Panel-2017 subjects made choices. Figure 6b shows the non-center budget lines for which the Only-2020 and the Panel-2020 subjects made choices. In 2017, subjects had to make 52 additional decisions (Figure 6a). Due to time constraints, we limited the number of additional choice situations in 2020 (Figure 6b). We predicted subjects' behavior for the non-center budget lines in Figure 6 based on a model that estimated distributional preference parameters exclusively on the basis of centered budget lines shown in Figure 1.

To describe the estimation of the distributional model, we introduce the following notation. For each budget line j, individual i chooses one out of seven possible allocations. Each allocation assigns a payoff to "self" and to "other", denoted by $w_{ij} = (w_{ij}^s, w_{ij}^o)$. Individual i's utility function V_i is given by

$$V_i\left(w_{ij}^s, w_{ij}^o\right) = w_{ij}^s - \alpha_i \max\left\{w_{ij}^o - w_{ij}^s, 0\right\} - \beta_i \max\left\{w_{ij}^s - w_{ij}^o, 0\right\},$$

which is identical to the functional form chosen by Fehr and Schmidt (1999). If the preference parameters are strictly positive ($\alpha_i > 0$ and $\beta_i > 0$), individual i is inequality averse, where α_i captures aversion against disadvantageous inequality and β_i the aversion against advantageous inequality. Note, however, that in the two-person case this functional form captures also altruistic utility functions like those of Charness and Rabin (2002) if we allow for individuals with $\alpha_i \leq 0$ and $\beta_i > 0$. In principle, the model also captures purely envious individuals if $\alpha_i > 0$ and $\beta_i < 0$. Pure self-interest is captured by $\alpha_i = 0$ and $\beta_i = 0$. We put no restrictions on the size or the sign of α_i and β_i in our empirical estimation, which ensures that the model nests all distributional preference types observed in the data.

We use discrete choice models assuming random utility, and also estimate an idiosyncratic error parameter $\zeta_i > 0$, in addition to the two behavioral parameters α_i and β_i. The value of an interpersonal allocation w_{ij} depends thus on three parameters summarized by $\theta_i' = (\alpha_i, \beta_i, \zeta_i)$. The discrete choice model yields, for each allocation, the choice probabilities:

$$\text{Prob}\left(V_i\left(w_{ij}\right) - V_i\left(w_{ij}'\right) > \varepsilon_{-ij} - \varepsilon_{ij} \right) = \frac{e^{\zeta_i V_i(w_{ij})}}{\sum_k e^{\zeta_i V_i(w_{kj})}},$$

where w_{ij}' indexes the allocation options *not* chosen by the individual i in choice situation j, ε_{ij} denotes the error term, and k indexes all the available choice options.

We estimate a Bayesian hierarchical model in which the (untransformed) individual-level parameters follow a multivariate normal distribution.[19] In addition, we use a diffuse prior, and draw from the posterior distribution using a Gibbs sampler.

For all our data sets, we estimate the model at different levels of aggregation. First, we estimate a representative agent model with only two parameters. Next, we estimate the model for each of the three preference types that we identified with the

---

[19]The procedure is described in detail in Allenby (1997) and Train (2001). The hierarchical Bayes approach also estimates mean and standard deviation of the population parameter distribution. Each individual is part of this distribution and is allowed to deviate from the sample mean. Thus, the model "disciplines" individual estimates in the sense that they may not depart "too strongly" from population-typical behavior. These models have the feature that individuals who show rather erratic behavior, or behavior departing strongly from typical behaviors, appear to be closer to the sample mean (this is called shrinkage).

DP-means algorithm, which yields 3 (types) × 2 = 6 estimated parameters. Finally, we estimate the preference parameters that capture individual-level heterogeneity, which leads to a set of parameters that is orders of magnitude larger.[20] Thus, the individual level model allows for a much richer empirical description of our data sets than the type-based or the representative agent model.

To compare the predictive ability of the different empirical models, we compute their hit rates. The hit rate summarizes how often the predictions exactly coincide with subject's actual choice. It ranges from 0 percent (when the model predicts all outcomes incorrectly) to 100 percent (when the model has perfect predictive accuracy). We have also computed the mean squared error (MSE) of each model which gives us the same conclusions as the hit rates. For this reason, we only report the hit rates below.

Table 6 below reports the hit rates of the different empirical models (columns 2-4), separately for each of the four data sets (the four rows). A striking feature of this table is that all models have a much better predictive ability than random choice. Chance would imply a hit rate of approximately 14 percent (1 allocation out of 7 possible allocations), but the hit rates reported for all the models estimated are all 3.5 to 5 times larger. Thus, all the models predict behavior much better than chance.

A second important result is that the three-type model has a considerably higher predictive ability than the representative agent model. The three-type model has a higher predictive accuracy for all data sets, and its predictive superiority is particularly pronounced for the Panel-2020 data set, where its hit rate is 18 percentage points higher, and the Only-2020 data set, where its hit rate is 15 percentage points higher.

Finally, the third key finding is that the hit rates of the three-type model are very similar to those of the individual-level model. In fact, there are even cases where the predictive accuracy of the type-based model is better than that of the individual-level model. For example, in the Panel-2017 sample, the individual-level model makes accurate predictions 70.4 percent of the time whereas the type-based model has a hit rate of 72.5 percent. Similar results hold for the Only-2017 data set, where the three-type model also fares better. It is, in our view, quite remarkable that a model with

---

[20]Specifically, the set of estimated preference parameters for the individual-level model varies between 348 x 2 = 796 (in the Only-2017 data set) and 916 x 2 = 1832 parameters (in the Only-2020 data set).

only three types (i.e., six estimated preference parameters in total) has a predictive accuracy that is as good as that of a model that has many more parameters.

Overall, these findings reveal a remarkable predictive ability of a model that is only based on three different preference types. Moving from one to three types leads to a substantial improvement in predictive power, but allowing for individual-level heterogeneity basically generates the same predictive ability as the three-type model.

Table 6: Comparing the out-of-sample predictive accuracy (hit rates) of the three-types model with an individual-level model and a representative agent model

Sample	Representative agent (k=1)	Three types (k=3)	Individual (k=N)
Panel-2017	0.653	**0.727**	0.703
Panel-2020	0.490	**0.672**	0.679
Only-2017	0.637	**0.715**	0.695
Only-2020	0.511	**0.665**	0.681

Note: The table displays the hit rates of the different empirical models (representative agent model, three-type model, individual-level model) for the different data sets (rows). The hit rates correspond to the share of choice situations for which the subjects' choices coincide exactly with the model's predictions. It ranges from 0 (when the model predicts all outcomes incorrectly) to 1 (when the model has a perfect predictive accuracy). Completely random choice behavior predicts a hit rate of 0.14.

4.2 Type-based versus machine learning-based predictions

Our three-type model is parsimonious and rests on a sound identification of the key motivational forces that govern redistributive behaviors. While the model is portable in the sense that it can be used to predict behavior in new choice situations like the non-center budget lines, how does it compare to a state-of-the-art machine learning method designed for high predictive ability? To answer this question, we train regularized gradient boosting trees (rGBT) on the same 12 center budget lines used for type identification and for the estimation of the structural distributional preference models discussed in Section 4.1. We then predict choices both within sample (i.e., for the center budget lines) and out of sample (i.e., for non-center budget lines) in the different data sets. Regularized GBT is widely used in computer science and has been demonstrated to outperform alternative machine learning models in various prediction scenarios because of its iterative error correction mechanism (Shwartz-Ziv

and Armon, 2022)[21], and it has also been successfully employed in a few economic applications (see, e.g., Chalfin et al., 2016; Einav et al., 2018; Deryugina et al., 2019).[22] Thus, we train regularized GBT on subjects' decisions on the estimation set and predict their decisions both for the estimation set (within-sample predictions) and for the choice situations in the prediction set (out-of-sample predictions).

When comparing the predictions of the machine learning approach and the structural economic approach, it is useful to precisely consider the information that is used to train or estimate the models. The structural economic approach with three preference types uses information about all individuals' choices on the 12 budget lines displayed in Figure 1 *and* each individual's type assignment, but restricts the estimated model to two preference parameters (α and β) for each type. Thus, the three-type model represents each individual of a particular type with that type's average (or "type-representative") preference parameters. This approach thus neglects the individuals' behavioral identity with the exception of their type-assignment. For the machine learning approach, we use budget line end points and individual identifiers as training inputs. In other words, we base our gradient boosting trees on a richer information base and thus expect rGBT to make better *within-sample* predictions than the structural three-type model.[23] Note, however, that this does not necessarily mean that the machine learning tool makes also better *out-of-sample* predictions because the non-center budget lines used for prediction differ from the centered budget lines, and rGBT does not capture the structural motivational forces underlying human behavior. rGBT might thus fare poorly when used to predict choices in novel situations. In contrast, the structural three-type model is applicable across domains and may have a better out-of-sample predictive ability than rGBT, to the extent that it captures the

---

[21]rGBT has also outperformed many alternative machine learning models in prediction competitions. See: https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions (retrieved on September 19, 2023).

[22]We employ the XGBoost algorithm (Chen and Guestrin, 2016), a regularized gradient boosting technique. In comparison to traditional gradient boosting, XGBoost introduces several innovations which benefit efficiency and mitigate overfitting. For our application, we set the learning rate, which scales the contribution of each tree, to 0.8. We specify the maximum tree depth to 10, and the maximum of boosting iterations to 100. We utilize a softmax objective function, so that the algorithm returns non-probabilistic predictions of allocation choices on each budget line.

[23]In principle, it is also possible to study the predictive power of gradient boosting trees in the case in which – like in the structural three-type model – it neglects individual identity information and only considers information on individuals' type (rGBT with type information). We discuss the results of this exercise in the next footnote.

fundamental motivational forces at work.

Table 7 compares the accuracy (hit rates) of the machine learning with the hit rates of the three-type model. Rows 1 to 4 correspond to the predictions for the center budget lines (i.e., the within-sample predictions) of the two approaches for the four data sets, while rows 5 to 8 correspond to the predictions for the non-center budget lines (i.e., the out-of-sample predictions). We deliberately display the within-sample predictive accuracy of gradient boosting trees to show that the machine learning tool is really good at what it has been designed for: maximizing predictive accuracy across structurally identical situations. In fact, as the first four rows in Table 7 show, the within-sample hit rates of gradient boosting trees vary between 97 to 99 percent, which is clearly better than the within-sample accuracy of the three-type model, which varies between 70 and 72 percent.[24]

What about predictive accuracy for the out-of-sample predictions? Here, the type-based model does substantially better than gradient boosting trees. While gradient boosting trees yields hit rates that are better than chance (recall that random choices would yield hit rates of about 0.14), they are all much lower than the hit rates obtained by the three-type model. In fact, the out-of-sample hit rate of the three-type model is roughly 2.3 times higher in all samples.

Taken together, these findings show that a parsimonious structural model of distributional preferences that captures the essence of individuals' fundamental distributional motives does a far better job in predicting behavior in novel situations compared to a state-of-the-art machine learning tool designed for high predictive ability. This finding underscores that, to be able to predict well across novel domains, understanding the structural motivational forces that shape behaviors is of paramount importance and of greater use than simply relying on a "theory-blind" machine learning tool, even when the empirical economic model is so parsimonious that it neglects many individual-level differences in preferences.[25]

---

[24]What happens if we examine the predictive power of gradient boosting trees when we neglect individual identity information and consider only information on individuals' type information (rGBT with type information). In this case, the within-sample predictive performance (HIT rate) of rGBT is only between 70 and 72%. Thus, individual identity information appears to be crucial for the within-sample predictive superiority of machine learning.

[25]Interestingly, we find that even the structural representative agent model makes considerably better out-of-sample predictions than rGBT. Due to space constraints we have currently omitted the detailed description of this finding from the paper.

Table 7: The out-of-sample predictive accuracy (hit rate) of the structural three-type-model and of gradient boosting trees

Sample	Prediction	Estimation set	Prediction set	Types (k=3)	Gradient boosting
Panel2017	within-sample	Center	Center	0.705	**0.995**
Panel2020	within-sample	Center	Center	0.708	**0.996**
Only2017	within-sample	Center	Center	0.700	**0.999**
Only2020	within-sample	Center	Center	0.724	**0.968**
Panel2017	out-of-sample	Center	Non-center	**0.727**	0.325
Panel2020	out-of-sample	Center	Non-center	**0.672**	0.292
Only2017	out-of-sample	Center	Non-center	**0.715**	0.289
Only2020	out-of-sample	Center	Non-center	**0.665**	0.279

Note: This table displays the hit rates of the structural three-type model (column 5) and of gradient boosting trees (column 6) in the different data sets (rows). Rows 1-4 correspond to within- sample predictions. Rows 5-8 correspond to the out-of-sample predictions. The hit rates reflect the share of choice situations where the subject's choice coincides exactly with the model's prediction. It ranges from 0 (when the model predicts all outcomes incorrectly) to 1 (when the model has a perfect predictive accuracy).

5   Summary and conclusion

Parsimony is widely considered to be a virtue in economic modelling. It rests on the idea that empirical and theoretical models should concentrate on the essential characteristics of the problem at hand. At the same time, parsimony is typically associated with neglecting details that may be important. So how much detail should we neglect, and how should we determine what is essential? We tackled this problem in the context of assessing the essential characteristics of preference heterogeneity in the domain of distributional preferences.

For this purpose, we used a Bayesian nonparametric clustering algorithm—DP-means—that makes the trade-off between parsimony and precision in the analysis of preference heterogeneity explicit and does not require any assumptions on the characteristics of distributional preferences. The empirical properties of the precision-parsimony frontier as well as the descriptive analysis of our four data sets provide strong hints that it may be possible to capture the essential characteristics of preference heterogeneity with three behavioral types.

The parsimony-precision frontier displays strongly decreasing precision gains from sacrificing parsimony by allowing more behavioral clusters and has a salient kink at three behavioral types in all four data sets—indicating small precision gains from more than three clusters and large precision losses from less than three clus-

ters. The descriptive analysis also hints at the existence of three large behavioral agglomerations but fails to to assign all individuals to a behavioral cluster.

We solve this problem by applying DP-means and assessing the behavioral interpretation of fewer and more than three clusters. This analysis shows that fewer clusters generate "dirty" preference agglomerations that merge very different preferences into the same clusters. Likewise, allowing for four or five clusters does not lead to empirically relevant and behaviorally meaningful new clusters while with three clusters we recover three clean, and fundamentally distinct, types with a clear behavioral interpretation—an inequality averse type, an altruistic type, and a predominantly selfish type. Remarkably, these three behavioral types emerge in all four data sets in roughly similar proportions, indicating a rather high stability of the type distribution both across samples and over time.

Finally, we show that relying on a three-type distribution to describe the behavioral heterogeneity does not necessarily mean that one has to sacrifice predictive accuracy, even when it comes to predicting the behavior of *individuals* out-of-sample. Indeed, if one uses the structurally estimated utility function of each behavioral type to predict the behavior of individuals out-of-sample, the predictive accuracy of this parsimonious model is equally good compared to the accuracy of a model that is based on individual-level estimates of utility functions. Moreover, the three-type model outperforms the representative agent model and makes far better out-of-sample predictions than a state-of-the-art machine learning tool. Thus, taken together, the three-type model not only gives us a parsimonious characterization of heterogeneity with a meaningful behavioral interpretation, but also appears to be a good tool for predicting individuals' behavior in novel situations.

While we gathered our data in Switzerland, an interesting question for future research would be to assess whether the population of other countries can also be parsimoniously characterized with a small number of distributional preference types, and whether the same or different types emerge. It would also be interesting to investigate whether the same types emerge in difference substrata of the same population such as, for example, in students or, more generally, in highly skilled strata of the population with tertiary education. Preliminary evidence suggests that a whole type—the inequality averse type—might not exist among students, and that higher education

generally tends to mitigate inequality aversion Epper ⓡ al. (2023). Finally, it could be interesting to examine whether the application of DP-means to the domain of risk and time preferences also leads to the identification of a parsimonious distribution of risk-taking and time discounting types.[26]

---

[26]Previous applications of mixture models to the domain of risk taking (Bruhin et al., 2010; Conte et al., 2011) provide hope that this may be possible. Under assumptions that restrict the feasible space of utility functionals, these papers identified an expected utility type and a rank-dependent utility type (resp. a cumulative prospect theory type).

References

Allenby, G (1997) "An introduction to hierarchical Bayesian modeling," in *Tutorial notes, Advanced Research Techniques Forum, American Marketing Association*.

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

Andrews, Isaiah, Drew Fudenberg, Lihua Lei, Annie Liang, and Chaofeng Wu (2022) "The Transfer Performance of Economic Models," *arXiv preprint arXiv:2202.04796*.

Bardsley, Nicholas and Peter G Moffatt (2007) "The experimetrics of public goods: Inferring motivations from contributions," *Theory and Decision*, 62, 161–193.

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

Bellemare, Charles, Sabine Kröger, and Arthur Van Soest (2008) "Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities," *Econometrica*, 76 (4), 815–839.

Bellemare, Charles and Bruce Shearer (2007) "Gift exchange within a firm: Evidence from a field experiment."

Bierbrauer, Felix and Nick Netzer (2016) "Mechanism design and intentions," *Journal of Economic Theory*, 163, 557–603.

Bierbrauer, Felix, Axel Ockenfels, Andreas Pollak, and Désirée Rückert (2017) "Robust mechanism design and social preferences," *Journal of Public Economics*, 149, 59–80.

Bolton, Gary E (1991) "A comparative model of bargaining: Theory and evidence," *The American Economic Review*, 1096–1136.

Bolton, Gary E and Axel Ockenfels (2000) "ERC: A theory of equity, reciprocity, and competition," *American Economic Review*, 90 (1), 166–193.

Breitmoser, Yves (2013) "Estimation of social preferences in generalized dictator games," *Economics Letters*, 121 (2), 192–197.

Breza, Emily, Supreet Kaur, and Nandita Krishnaswamy (2019) "Coordination without organization: Collective labor supply in decentralized spot markets,"Technical report, National Bureau of Economic Research.

Breza, Emily, Supreet Kaur, and Yogita Shamdasani (2018) "The morale effects of pay inequality," *The Quarterly Journal of Economics*, 133 (2), 611–663.

———— (2021) "Labor rationing," *American Economic Review*, 111 (10), 3184–3224.

Bruhin, Adrian, Helga Fehr-Duda, and Thomas Epper (2010) "Risk and rationality: Uncovering heterogeneity in probability distortion," *Econometrica*, 78 (4), 1375–1412.

Bruhin, Adrian, Ernst Fehr, and Daniel Schunk (2018) "The many faces of human sociality: Uncovering the distribution and stability of social preferences," *Journal of the European Economic Association*, 17 (4), 1025–1069.

Burghart, Daniel R, Thomas Epper, and Ernst Fehr (2020) "The uncertainty triangle–Uncovering heterogeneity in attitudes towards uncertainty," *Journal of Risk and Uncertainty*, 60 (2), 125–156.

Buschena, David and David Zilberman (2000) "Generalized expected utility, heteroscedastic error, and path dependence in risky choice," *Journal of Risk and Uncertainty*, 20 (1), 67–88.

Camerer, Colin F (2011) *Behavioral game theory: Experiments in strategic interaction*: Princeton University Press.

Camerer, Colin F and George Loewenstein (1993) "Information, fairness, and efficiency in bargaining."

Camerer, Colin F, Gideon Nave, and Alec Smith (2019) "Dynamic unstructured bargaining with private information: theory, experiment, and outcome prediction via machine learning," *Management Science*, 65 (4), 1867–1890.

Camerer, Colin and Richard H Thaler (1995) "Anomalies: Ultimatums, dictators and manners," *Journal of Economic Perspectives*, 9 (2), 209–219.

Cappelen, Alexander W, Astri Drange Hole, Erik Ø Sørensen, and Bertil Tungodden (2007) "The Pluralism of Fairness Ideals: An Experimental Approach," *American Economic Review*, 97 (3), 818–827.

Carlsson, Fredrik, Olof Johansson-Stenman, and Pham Khanh Nam (2014) "Social preferences are stable over long periods of time," *Journal of public economics*, 117, 104–114.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan (2016) "Productivity and selection of human capital with machine learning," *American Economic Review*, 106 (5), 124–127.

Charness, Gary (2000) "Responsibility and effort in an experimental labor market," *Journal of Economic Behavior & Organization*, 42 (3), 375–384.

Charness, Gary and Matthew Rabin (2002) "Understanding social preferences with simple tests," *Quarterly Journal of Economics*, 817–869.

Chen, Tianqi and Carlos Guestrin (2016) "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794.

Chuang, Yating and Laura Schechter (2015) "Stability of experimental and survey measures of risk, time, and social preferences: A review and some new results," *Journal of development economics*, 117, 151–170.

Comiter, Marcus Zachary, Miriam Cha, HT Kung, and Surat Teerapittayanon (2016) "Lambda means clustering: automatic parameter search and distributed computing implementation," *2016 23rd international conference on pattern recognition (ICPR)*, 2331–2337.

Conte, Anna, John D Hey, and Peter G Moffatt (2011) "Mixture models of choice under risk," *Journal of Econometrics*, 162 (1), 79–88.

Conte, Anna and M Vittoria Levati (2014) "Use of data on planned contributions and stated beliefs in the measurement of social preferences," *Theory and Decision*, 76, 201–223.

Conte, Anna and Peter G. Moffatt (2012) "The econometric modelling of social preferences," *Theory and Decision*, 76 (1), 119–145.

Deryugina, Tatyana, Garth Heutel, Nolan H Miller, David Molitor, and Julian Reif (2019) "The mortality and medical costs of air pollution: Evidence from changes in wind direction," *American Economic Review*, 109 (12), 4178–4219.

Dur, Robert (2009) "Gift exchange in the workplace: Money or attention?" *Journal of the European Economic Association*, 7 (2-3), 550–560.

Durante, Ruben, Louis Putterman, and Joël Van der Weele (2014) "Preferences for redistribution and perception of fairness: An experimental study," *Journal of the European Economic Association*, 12 (4), 1059–1086.

Einav, Liran, Amy Finkelstein, Sendhil Mullainathan, and Ziad Obermeyer (2018) "Predictive modeling of US health care spending in late life," *Science*, 360 (6396), 1462–1465.

Epper, Thomas ⓡ Julien Senn ⓡ Ernst Fehr (2023) "The missing type: where are the inequality averse (students)?" *Working paper series, Department of Economics, University of Zurich* (435).

Fehr, Ernst ⓡ Thomas Epper ⓡ Julien Senn (2021a) "Other-regarding preferences and redistributive politics," *Working paper series, Department of Economics, University of Zurich* (339).

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl (1993) "Does fairness prevent market clearing? An experimental investigation," *The quarterly journal of economics*, 108 (2), 437–459.

Fehr, Ernst, Michael Powell, and Tom Wilkening (2021b) "Behavioral constraints on the design of subgame-perfect implementation mechanisms," *American Economic Review*, 111 (4), 1055–1091.

Fehr, Ernst and K Schmidt (1999) "A Theory Of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114 (3), 817–868.

Fisman, Raymond, Pamela Jakiela, and Shachar Kariv (2017) "Distributional preferences and political behavior," *Journal of Public Economics*, 155, 1–10.

Fisman, Raymond, Pamela Jakiela, Shachar Kariv, and Daniel Markovits (2015) "The distributional preferences of an elite," *Science*, 349 (6254).

Fisman, Raymond, Pamela Jakiela, Shachar Kariv, and Silvia Vannutelli (2023) "The distributional preferences of Americans, 2013–2016," *Experimental Economics*, 1–22.

Fisman, Raymond, Shachar Kariv, and Daniel Markovits (2007) "Individual preferences for giving," *American Economic Review*, 97 (5), 1858–1876.

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan (2022) "Measuring the completeness of economic models," *Journal of Political Economy*, 130 (4), 956–990.

Fudenberg, Drew and Indira Puri (2021) "Evaluating and extending theories of choice under risk," *Working paper, MIT Economics*.

Gächter, Simon, Daniele Nosenzo, and Martin Sefton (2013) "Peer effects in pro-social behavior: Social norms or social preferences?" *Journal of the European Economic Association*, 11 (3), 548–573.

Gächter, Simon and Christian Thöni (2005) "Social learning and voluntary cooperation among like-minded people," *Journal of the European Economic Association*, 3 (2-3), 303–314.

——— (2010) "Social comparison and performance: Experimental evidence on the fair wage–effort hypothesis," *Journal of Economic Behavior & Organization*, 76 (3), 531–543.

Iriberri, Nagore and Pedro Rey-Biel (2011) "The role of role uncertainty in modified dictator games," *Experimental Economics*, 14, 160–180.

——— (2013) "Elicited beliefs and social information in modified dictator games: What do dictators believe other dictators do?" *Quantitative Economics*, 4 (3), 515–547.

Kerschbamer, Rudolf and D Müller (2020) "Social preferences and political attitudes: An online experiment on a large heterogeneous sample," *Journal of Public Economics*, 182.

Kirchsteiger, Georg (1994) "The role of envy in ultimatum games," *Journal of economic behavior & organization*, 25 (3), 373–389.

Kube, Sebastian, Michel André Maréchal, and Clemens Puppe (2012) "The currency of reciprocity: Gift exchange in the workplace," *American Economic Review*, 102 (4), 1644–1662.

Kulis, Brian and Michael I. Jordan (2012) "Revisiting k-means: New Algorithms via Bayesian Nonparametrics," *Proceedings of the 29th International Conference of Machine Learning*.

Plonsky, Ori, Reut Apel, Eyal Ert et al. (2019) "Predicting human decisions with behavioral theories and machine learning," *arXiv preprint arXiv:1904.06866*.

Plonsky, Ori, Ido Erev, Tamir Hazan, and Moshe Tennenholtz (2017) "Psychological forest: Predicting human behavior," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 31.

Schmidt, Klaus M and Axel Ockenfels (2021) "Focusing climate negotiations on a uniform common commitment can promote cooperation," *Proceedings of the National Academy of Sciences*, 118 (11), e2013070118.

Shwartz-Ziv, Ravid and Amitai Armon (2022) "Tabular data: Deep learning is not all you need," *Information Fusion*, 81, 84–90.

Train, Kenneth (2001) "A comparison of hierarchical Bayes and maximum simulated likelihood for mixed logit," *University of California, Berkeley*, 1–13.

Tyran, Jean-Robert and Rupert Sausgruber (2006) "A little fairness may induce a lot of redistribution in democracy," *European Economic Review*, 50 (2), 469–485.

ONLINE APPENDIX

A Experimental task and population sample

A.1 Details on choice situations

Table A.1 describes the 12 choice situations presented in Figure 1. These are the choice situations that we use to identify subjects' other-regarding preferences, both for the clustering (Section 3.2) and for the structural estimation of a two-parameter model of distributional preferences (Section 4.1). The definition of the different variables in the Table is as follows. 'choiceId' is the unique identifier for each choice situation. (own1, other1) represents the payoff combination at the lower end of the budget line (in points). (own2, other2) represents the payoff combination at the upper end of the budget line (in points). 'bundle' indicates to which bundle the respective choice situation belongs to (centered vs. non-center), and 'slope' denotes the slope of the budget line in the "own payoff – other payoff" space.

Table A.1: Estimation set (centered budget lines)

choiceId	own1	own2	other1	other2	bundle	slope
1	450	1050	750	750	center	0.0
2	500	1000	800	700	center	-0.2
3	550	950	850	650	center	-0.5
4	600	900	900	600	center	-1.0
5	650	850	950	550	center	-2.0
6	700	800	1000	500	center	-5.0
7	750	750	1050	450	center	-Inf
8	700	800	500	1000	center	5.0
9	650	850	550	950	center	2.0
10	600	900	600	900	center	1.0
11	550	950	650	850	center	0.5
12	500	1000	700	800	center	0.2

Tables A.2 and A.3 describe the choice situations used for out-of-sample predictions in the 2017 and the 2020 waves, respectively.

Table A.2: Prediction set (2017 wave, non-centered budget lines)

choiceId	own1	own2	other1	other2	bundle	slope
15	300	600	900	600	non-center	-1.0
16	600	900	1200	900	non-center	-1.0
17	300	900	900	900	non-center	0.0
18	350	850	950	850	non-center	-0.2
19	400	800	1000	800	non-center	-0.5
20	450	750	1050	750	non-center	-1.0
21	500	700	1100	700	non-center	-2.0
22	550	650	1150	650	non-center	-5.0
23	600	600	1200	600	non-center	-Inf
24	550	650	650	1150	non-center	5.0
25	500	700	700	1100	non-center	2.0
26	450	750	750	1050	non-center	1.0
27	400	800	800	1000	non-center	0.5
28	350	850	850	950	non-center	0.2
29	600	900	600	300	non-center	-1.0
30	900	1200	900	600	non-center	-1.0
31	600	1200	600	600	non-center	0.0
32	650	1150	650	550	non-center	-0.2
33	700	1100	700	500	non-center	-0.5
34	750	1050	750	450	non-center	-1.0
35	800	1000	800	400	non-center	-2.0
36	850	950	850	350	non-center	-5.0
37	900	900	900	300	non-center	-Inf
38	850	950	350	850	non-center	5.0
39	800	1000	400	800	non-center	2.0
40	750	1050	450	750	non-center	1.0
41	700	1100	500	700	non-center	0.5
42	650	1150	550	650	non-center	0.2
43	250	850	950	950	non-center	0.0
44	300	800	1000	900	non-center	-0.2
45	350	750	1050	850	non-center	-0.5
46	400	700	1100	800	non-center	-1.0
47	450	650	1150	750	non-center	-2.0
48	500	600	1200	700	non-center	-5.0
49	550	550	1250	650	non-center	-Inf
50	600	700	600	1100	non-center	5.0
51	600	800	600	1000	non-center	2.0
52	500	900	700	900	non-center	0.5
53	400	900	800	900	non-center	0.2
54	650	1250	550	550	non-center	0.0
55	700	1200	600	500	non-center	-0.2
56	750	1150	650	450	non-center	-0.5
57	800	1100	700	400	non-center	-1.0
58	850	1050	750	350	non-center	-2.0
59	900	1000	800	300	non-center	-5.0
60	950	950	850	250	non-center	-Inf
61	600	1100	600	700	non-center	0.2
62	600	1000	600	800	non-center	0.5
63	700	900	500	900	non-center	2.0
64	800	900	400	900	non-center	5.0

Table A.3: Prediction set (2020 wave, non-centered budget lines)

choiceId	own1	own2	other1	other2	bundle	slope
13	700	767	800	1133	non-center	5.0
14	800	950	700	250	non-center	-3.0
15	700	800	800	1100	non-center	3.0
16	800	1100	700	400	non-center	-1.0
17	700	820	800	1080	non-center	2.3
18	800	1175	700	575	non-center	-0.3
19	700	718	800	1185	non-center	21.4
20	800	830	700	130	non-center	-19.0

A.2 Descriptive statistics

Table A.4: Comparison of sample population with the Swiss population

	Panel-2017	Panel-2020	Only-2017	Only-2020	Population
Male	0.56	0.56	0.52	0.48	0.48
Age (mean)	48.13	51.20	44.27	43.26	51.08
Education : Obligatory school	0.03	0.03	0.05	0.02	0.11
Education : Vocational training	0.34	0.35	0.41	0.34	0.42
Education : High school	0.15	0.10	0.11	0.13	0.10
Education : University	0.37	0.41	0.30	0.39	0.35
Education : Other	0.10	0.11	0.12	0.11	-
Income bracket : $\leq$ CHF 4000	0.26	0.28	0.24	0.37	0.28
Income bracket : CHF 4001-6000	0.17	0.23	0.24	0.21	0.26
Income bracket : CHF 6001-8000	0.20	0.20	0.20	0.18	0.22
Income bracket : CHF 8001-10000	0.15	0.14	0.13	0.10	0.12
Income bracket : CHF 10001-15000	0.10	0.09	0.08	0.05	0.09
Income bracket : $\geq$ CHF 15000	0.02	0.01	0.01	0.03	0.03
Income bracket : NA	0.10	0.05	0.09	0.05	-
Unemployed	0.03	0.02	0.04	0.03	0.03
N	467	467	348	916	

Notes: The table displays descriptive statistics (mean) for the main socio-demographics of the main sample and for the Swiss population. The population data were obtained from the Swiss Federal Bureau of Statistics (2018) and are restricted to the adult Swiss population (i.e. individuals holding a swiss passport who are at least 18 years old).

B Alternative number of clusters

B.1 Allowing for four clusters

In this Appendix, we display—for each sample—the transition matrices that document how individuals assignment to clusters varies when we increase precision such that the number of clusters allowed to emerge *increases* from $k = 3$ to $k = 4$. In all the samples, the vast majority of individuals remain assigned to the same behavioral cluster. For example, in the Panel-2017 data set (Table A.5), 224 individuals out of the 225 assigned to the inequality averse cluster when $k = 3$ remain assigned to the same cluster when $k = 4$. Similarly, the whole 196 individuals assigned to the altruistic cluster and the 46 assigned to the predominantly selfish cluster remain assigned to the same clusters. In the remaining data sets, we observe similar patterns: almost all individuals remain assigned to the same behavioral cluster, and the fourth cluster remains populated by very few individuals' whose behavior is hard to interpret (See Tables A.6 and A.7). We depicted the transition matrix for the Only-2020 sample directly in the main text (Table 3). As we discuss there, the inequality averse cluster splits into two sub-clusters in this sample. This interpretation is further supported by Figure A.1, which shows that the two clusters display a behavior that is consistent with inequality aversion. The other two clusters behave in a way consistent with selfishness (lower right panel) and altruism (lower left panel). Thus, in this sample too, allowing for four clusters does not reveal a new, distinct, and empirically relevant behavioral type.

5

Figure A.1: The distribution of choices for positively and negatively sloped budget lines in each cluster in the Only-2020 dataset when $k = 4$.
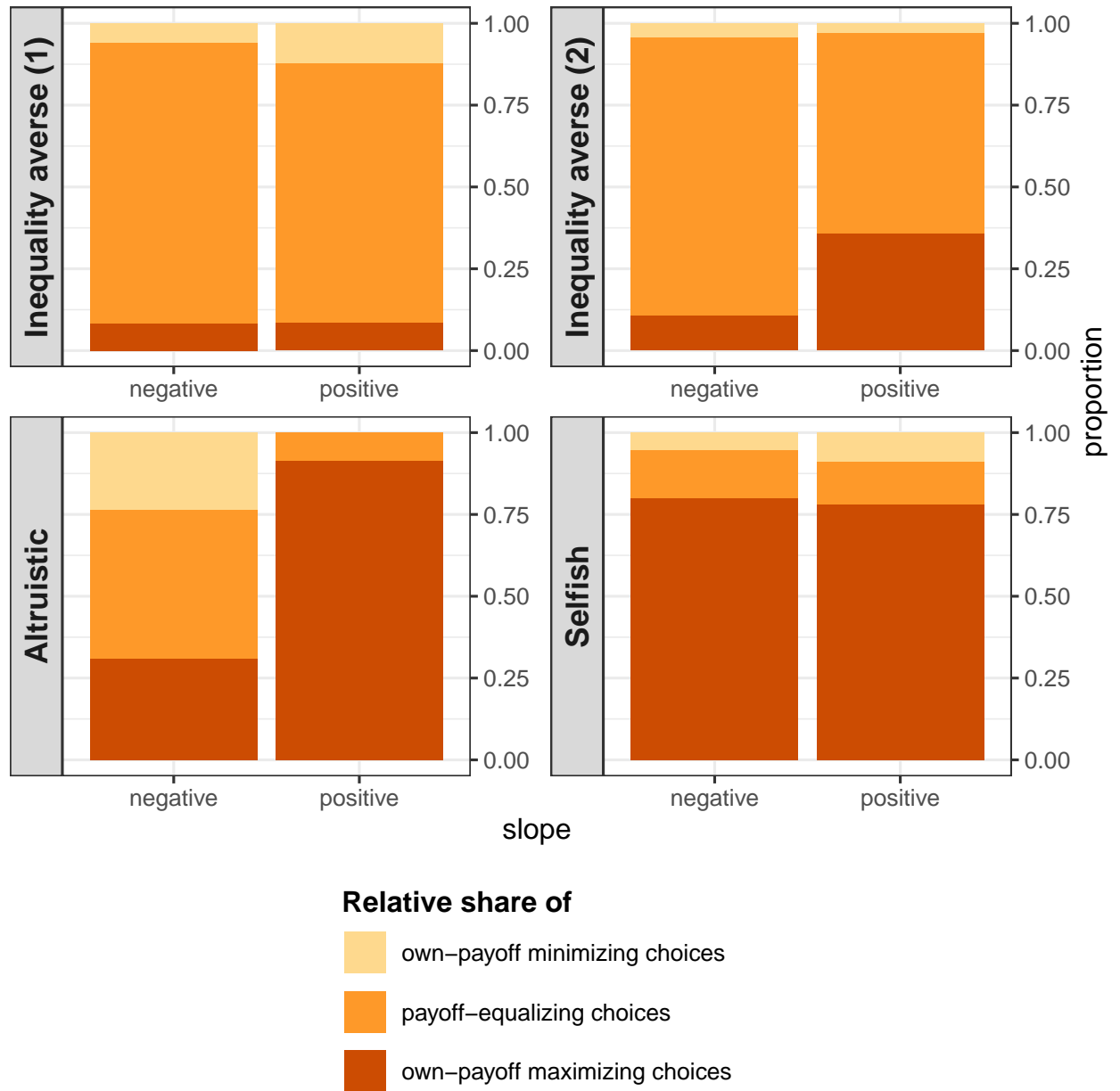
Table A.5: Transition of individuals between types in the Panel-2017 data set

		k = 4 clusters				
		Inequality averse	Altruistic	Predominantly Selfish	Non Interpretable	Total (%)
k=3	**Inequality averse**	224	0	0	1	225 (48.18%)
clusters	**Altruistic**	0	196	0	0	196 (41.97%)
	Predominantly selfish	0	0	46	0	46 (9.85%)
	Total (%)	224 (47.97%)	196 (41.97%)	46 (9.85%)	1 (0.21%)	467 (100%)

Table A.6: Transition of individuals between types in the Panel-2020 data set

		k = 4 clusters				
		Inequality averse	Altruistic	Predominantly selfish	Non Interpretable	Total (%)
k=3	**Inequality averse**	202	0	0	9	211 (45.18%)
clusters	**Altruistic**	0	181	0	0	181 (38.76%)
	Predominantly selfish	4	0	71	0	75 (16.06%)
	Total (%)	206 (44.11%)	181 (38.76%)	71 (15.20%)	9 (1.93%)	467 (100%)

Table A.7: Transition of individuals between types in the Only-2017 data set

		k = 4 clusters				
		Inequality averse	Altruistic	Predominantly Selfish	Non Interpretable	Total (%)
k=3	**Inequality averse**	186	0	0	1	187 (53.74%)
clusters	**Altruistic**	1	102	0	0	103 (29.59%)
	Predominantly selfish	0	0	58	0	58 (16.67%)
	Total (%)	187 (53.74%)	102 (29.31%)	58 (16.47%)	1 (0.29%)	348 (100%)

B.2 Allowing for five clusters

In this Appendix, we display—for each sample—the transition matrices that document how individuals' assignment to clusters varies when we increase precision such that the number of clusters allowed to emerge *increases* from $k = 3$ to $k = 5$. As we show below, we find again that the majority of individuals are assigned to an inequality averse cluster (IA), a smaller yet large group of individuals is assigned to an altruistic cluster, and the remaining individuals are assigned to a predominantly selfish cluster. However, these three behavioral types are now distributed in some data sets over a larger number of clusters while in other data sets the fourth and the fifth cluster basically remain unpopulated. This means that with five clusters the stability of the behavioral interpretation of the different clusters across data sets and time is completely lost. Table A.8 below, which depicts the respective distributions of behavioral clusters, illustrates this in detail.

Table A.8: Type distribution when allowing for five types

	Panel-2017	**Panel-2020**	**Only-2017**	**Only-2020**
Cluster 1	Altruistic (38.33)	IA (42.61)	IA (53.45)	IA-1 (26.53)
Cluster 2	IA-1 (33.19)	Altruistic (38.76)	Altruistic (29.31)	Selfish (21.18)
Cluster 3	IA-2 (18.63)	Selfish (15.20)	Selfish (16.67)	IA-2 (19.43)
Cluster 4	Selfish (9.64)	- (2.78)	- (0.29)	Altruistic-1 (18.78)
Cluster 5	- (0.21)	- (0.64)	- (0.29)	Altruistic-2 (14.08)

Note: The table displays the distribution of individuals to five clusters (in percent) in each of our four data sets. IA indicates inequality averse behavior, and IA-1 (IA-2) indicate the two clusters of inequality averse subjects. Altruistic-1 (Altruistic-2)indicate the two clusters of altruistic subjects. Selfish indicates the primarily self-interested behavioral type.

In two data sets (Panel-2020 and Only-2017), only three clusters are populated (one cluster for each behavioral type) and the remaining two clusters (Cluster 4 and Cluster 5) remain essentially unpopulated. In the Panel-2017 data set, the same three types emerge but the inequality averse type is divided up into two clusters (subtypes, IA-1 and IA-2) with an identical behavioral interpretation, as evidenced in the transition matrix (Table A.9). The fifth cluster is populated by a single individual. Finally, in the Only-2020 data set, all five clusters are populated, but this is explained by the fact that the inequality averse type and the altruistic type are both split into two clusters, as evidenced in Table A.10. Altogether, these results suggest that increasing precision such that 5 types are allowed to emerge does not bring fundamentally new insights compared to the preferred clustering with $k = 3$, and it undermines the stability of the behavioral interpretation of the clusters across data sets and across time.

Table A.9: Transition of individuals between types in the Panel-2017 data set

		k = 5 clusters					
		Inequality averse (1)	Inequality averse (2)	Altruistic	Predominantly Selfish	Non Interpretable	Total (%)
k=3 clusters	Inequality averse	154	70	0	0	1	225 (48.18%)
	Altruistic	0	17	179	0	0	196 (41.97%)
	Predominantly selfish	1	0	0	45	0	46 (9.85%)
	Total (%)	155 (33.19%)	87 (18.63%)	179 (38.33%)	45 (9.64%)	1 (0.21%)	467 (100%)

Table A.10: Transition of individuals between types in the Only-2020 data set

		k = 5 clusters					
		Inequality averse (1)	Inequality averse (2)	Altruistic (1)	Altruistic (2)	Predominantly Selfish	Total (%)
k=3 clusters	Inequality averse	175	230	12	0	0	417 (45.52%)
	Altruistic	0	1	155	123	0	279 (30.46%)
	Predominantly selfish	3	12	5	6	194	220 (24.02%)
	Total (%)	178 (19.43%)	243 (26.53%)	172 (18.78%)	129 (14.08%)	194 (21.18%)	916 (100%)

Table A.11: Transition of individuals between types in the Only-2017 data set

		k = 5 clusters					
		Inequality averse	Altruistic	Predominantly Selfish	Non Interpretable	Non Interpretable	Total (%)
k=3 clusters	Inequality averse	185	0	0	1	1	187 (53.74%)
	Altruistic	1	102	0	0	0	103 (29.60%)
	Predominantly selfish	0	0	58	0	0	58 (16.67%)
	Total (%)	186 (53.45%)	102 (29.31%)	58 (16.67%)	1 (0.29%)	1 (0.29%)	348 (100%)

Table A.12: Transition of individuals between types in the Panel-2020 data set

		k = 5 clusters					
		Inequality averse	Altruistic	Predominantly Selfish	Non Interpretable	Non Interpretable	Total (%)
k=3 clusters	Inequality averse	195	0	0	13	3	211 (45.18%)
	Altruistic	0	181	0	0	0	181 (38.76%)
	Predominantly selfish	4	0	71	0	0	75 (16.06%)
	Total (%)	199 (42.61%)	181 (38.76%)	71 (15.2%)	13 (2.78%)	3 (0.64%)	467 (100%)

B.3   Allowing for only two clusters

In this Appendix, we display the transition matrices that document how individuals assignment to types varies when we decrease precision (increase parsimony) such that the number of types allowed to emerge *decreases* from $k = 3$ to $k = 2$. We depict the results for the Panel-2020, the Only-2017 and the Only-2020 samples (evidence for the Panel-2017 is discussed in the main text). These tables show that, in all samples, restricting the number of types to two systematically leads to the disappearance of an important behavioral type and to the merging of incompatible preference types (denoted "MIP" in the tables) into one type, which undermines the behavioral interpretation of the different clusters. For example, in the Panel-2020 data set (Table A.13), the altruistic and the predominantly selfish types when $k = 3$ are merged together into a single, uninterpretable cluster when $k = 2$.

Table A.13: Transition of individuals between types in the Panel-2020 data set

		k = 3 clusters			
		Inequality averse	Altruistic	Predominantly Selfish	Total (%)
k=2	Inequality averse	203	6	7	216 (46.25%)
clusters	MIP	8	175	68	251 (53.75%)
	Total (%)	211 (45.18%)	181 (38.76%)	75 (16.06%)	467 (100%)

Table A.14: Transition of individuals between types in the Only-2017 data set

		k = 3 clusters			
		Inequality averse	Altruistic	Predominantly Selfish	Total (%)
k=2	MIP	187	100	1	288 (82.76%)
clusters	Predominantly Selfish	0	3	57	60 (17.24%)
	Total (%)	187 (53.73%)	103 (29.6%)	58 (16.67%)	348 (100%)

Table A.15: Transition of individuals between types in the Only-2020 data set

		k = 3 clusters			
		Inequality averse	Altruistic	Predominantly Selfish	Total (%)
k=2	MIP	412	3	80	495 (54.04%)
clusters	MIP	5	276	140	421 (45.96%)
	Total (%)	417 (45.52%)	279 (30.46%)	220 (24.02%)	916 (100%)

C Recoverability of Preference Types

In this appendix, we demonstrate the ability of the DP means algorithm to recover preference types from data. To that end, we simulate individual choices from the family of constant elasticity of substitution (CES) preferences which has been widely used in the literature (see, for example, Fisman, Jakiela and Kariv, 2017). A decision maker with CES preferences maximizes the utility function[27]

$$V\left(w^{(s)}, w^{(o)}\right) = v^{-1}\left(\kappa v\left(w^{(s)}\right) + (1-\kappa)v\left(w^{(o)}\right)\right) + \varepsilon,$$

where $w^{(s)}$ and $w^{(o)}$ denote the money allocation to self and the other person, respectively. v is a power function with parameter ρ

$$v(x) = \begin{cases} x^\rho & \text{if } \rho > 0 \\ \ln(x) & \text{if } \rho = 0 \\ -x^\rho & \text{if } \rho < 0 \end{cases},$$

and v^{-1} its inverse. The model has three parameter: κ, which governs the degree of fair-mindedness; ρ, which captures the equality-efficiency tradoffs; and ζ, which is the standard deviation of the errors controlling the deviations from deterministic maximization of the CES function (i.e. the utility shocks ε).[28]

We employ a hierarchical simulation exercise and generate 1000 individuals' allocation choices for three distinct preference types:

1. **An egalitarian altruist** (Rawlsian) type with mean parameters $\kappa = 0.5$ and a very low $\rho = -100$. This type approaches a maximin type with $V(\cdot) = \min\left\{w^{(s)}, w^{(o)}\right\}$ when $\rho \to -\infty$. This type's indifference curves are L-shaped.

2. **A strong altruist** type with mean parameters $\kappa = 0$ and $\rho = 0$ for which the power function converges to the logarithm, such that $V(\cdot) = \ln w^{(o)}$. This type's indifference curves are horizontal lines.

3. **A selfish** type with mean parameters $\kappa = 1$ and $\rho = 0$, such that $V(\cdot) = \ln w^{(s)}$. This type's indifference curves are vertical lines.

We allow for heterogeneity within preference types. Specifically, we assume that the preference parameter ρ and the logit-transformed κ are normally distributed with the above means and standard deviation 0.2. The logit transformation of the latter parameter ensures that the simulated κs lie within the unit interval.

We generate simulated data for a population consisting of 400 egalitarian altruists (40%), 400 strong altruists (40%), and 200 selfish individuals (20%). We consider three levels of noise: **low noise** ($\zeta = 0.01$), **medium noise** ($\zeta = 30$) and **high noise** ($\zeta = 60$). The expectation is that simulated types generated with more noise are less discernable from each other.

---

[27]For the sake of brevity, we omit individual-, choice-situation- and alternative-specific indices.

[28]Thus, κ largely determines the slope of the indifference curve, whereas ρ determines its curvature.

We restrict our attention to a level of precision that yields three clusters. Once again, we cluster based on the *center bundle* choice only. As mentioned in the main text, the DP-means algorithm is not aware of the interpretation of the emerging clusters. The cluster labels have to be assigned by the researcher after a careful inspection of the behavioral characteristics of each identified cluster, provided that the clusters have a clear behavioral interpretation.

In all scenarios, the DP-means recovers all three behavioral types. However, as the level of noise increases, each types' characteristic behavior displays larger variance, as shown in Figure A.2.

Figure A.2: The distribution of choices by recovered preference type and different noise levels.