Pareto Damaging Behaviors*

Raymond Fisman† Columbia University
Shachar Kariv‡ UC Berkeley
Daniel Markovits§ Yale University

July 20, 2005

Abstract

This paper reports a rigorous experimental test of Pareto-damaging behaviors. We introduce a new graphical representation of dictator games with step-shaped sets of feasible payoffs to persons *self* and *other* on which strongly Pareto efficient allocations involve substantial inequality. The non-convexity and sharp nonlinearity of the Pareto frontier allow us systematically to classify Pareto-damaging allocations: as *self*-damaging or *other*-damaging and as inequality-increasing or inequality-decreasing. We find that *self* and *other* Pareto-damaging behaviors occur frequently even in circumstances – dictator games – that do not implicate reciprocity or strategic interaction. We also find patterns in this behavior, most notably that behavior that Pareto damages *self* always reduces inequality whereas behavior that Pareto damages *other* usually increases inequality. (*JEL*: C79, C91, D64)

1 Introduction

The Pareto principle pervades economic conceptions of rationality. A large and growing body of laboratory evidence suggests, however, that participants in ex-

*This research was supported by the Experimental Social Science Laboratory (X-Lab) at the University of California, Berkeley. We thank Gary Charness for instructive conversations and detailed comments on an earlier draft. We are also grateful to Jim Andreoni, Stefano DellaVigna, Botond Koszegi, David Levine, Matthew Rabin, Andrew Schotter and José Silva for helpful discussions. We thank Brenda Naputi and Lawrence Sweet from the X-Lab for their valuable assistance, and Roi Zemmer for writing the experimental computer program. For financial support, Fisman thanks the Columbia University Graduate School of Business; Kariv acknowledges University of California, Berkeley; Markovits thanks Yale Law School and Deans Anthony Kronman and Harold Koh.

†Graduate School of Business, Columbia University, Uris 823, New York, NY 10027 (E-mail: rf250@columbia.edu, URL: http://www-1.gsb.columbia.edu/faculty/rfisman/).

‡Department of Economics, University of California, Berkeley, 549 Evans Hall # 3880, Berkeley, CA 94720 (E-mail: kariv@berkeley.edu, URL: http://socrates.berkeley.edu/~kariv/).

§Yale Law School, P.O. Box 208215, New Haven, CT 06520. (E-mail: daniel.markovits@yale.edu, URL: http://www.law.yale.edu/outside/html/faculty/ntuser93/profile.htm)

perimental dictator, ultimatum, and trust games often violate the Pareto principle, that is, often display Pareto-damaging behaviors. This challenges many of the models of other-regarding, or social, preferences that have been used in order to explain behavior in these games. In particular, *social welfare* models, which propose that persons pursue an aggregate of their own and others' payoffs, cannot account for Pareto-damaging behavior at all. And even *difference aversion* models, which allow persons to care about differences between their own and others' payoffs, can account for Pareto-damaging behavior only insofar as it reduces inequality. Experimental subjects, however, exhibit Pareto-damaging behavior that both decreases and increases inequality.

In this paper, we report on a laboratory experiment that allows for a rigorous test of Pareto-damaging behaviors and enables us better to distinguish among a range of competing models of social preferences that incorporate such behavior. We restrict attention to a dictator game and ignore the complications of strategic behavior and reciprocity motivations in response games in order to focus on Pareto-damaging behavior motivated by purely distributional preferences. Non-strategic behavior is simpler to analyze and is, moreover, adequate for comparing several prominent models of social preferences. Furthermore, purely distributive Pareto-damaging behaviors can arise in a wide variety of very common social and economic circumstances in the real world.

We use a new experimental design - that employs graphical representations of modified dictator games - in which each subject faces a large and rich menu of *step-shaped dictator sets* representing the feasible monetary payoffs to person $self$ and *other*. An example of one such step-shaped set Π is illustrated in Figure 1. Each point $\pi = (\pi_o, \pi_s)$ corresponds to the payoffs to persons *other* and $self$, respectively, and there are only two strictly Pareto efficient allocations: π^s maximizes the payoff for $self$ and π^o maximizes the payoff for *other*.

[Figure 1 here]

The step-shaped set enables us to distinguish effectively among several types of Pareto-damaging behaviors. Most importantly, the non-convexity and sharp nonlinearity of the Pareto frontier means that the dictator always faces choices with an extreme *relative price of giving*. In this context, either $self$ or *other* must be made strictly worse off in order to create greater equality or greater inequality. Hence, in contrast to typical split-the-pie dictator games, reducing or increasing differences in payoffs will generally involve Pareto-damaging behavior.

In particular, for each subject we distinguish between decisions that are inequality-increasing and inequality-decreasing, and we separate decisions that Pareto damage $self$ and that Pareto damage *other*. This allows us to differentiate among various prototypical preferences – competitive, self-interested, lexicographic for $self$ over *other*, difference averse, and social welfare. Moreover, the graphical representation of the feasible sets enables us to collect many more observations per subject than has heretofore been possible and therefore to analyze preferences at the level of the individual subject. The graphical representation also enables us to avoid emphasizing any particular allocation and,

critically, does not force subjects into discrete choices that suggest extreme prototypical preference types.

Our main findings can be summarized as follows. First, very few allocations Pareto damage both *self* and *other* by violating weak Pareto efficiency. Second, almost all strictly Pareto efficient allocations maximize the payoff for *self*. Third, nearly a third of all allocations Pareto damage either *self* or *other* (but not both) and therefore are only weakly Pareto efficient. Fourth, nearly three-quarters of Pareto-damaging allocations are *other* damaging. Fifth, and perhaps most importantly, all *self* Pareto-damaging allocations decrease inequality, while most *other* Pareto-damaging allocations increase inequality.

Finally, because of our rich data set, we are able to analyze preferences at the individual level. The majority of subjects have cleanly classifiable preferences, ranging from competitive to selfish to lexicographic to difference averse to social welfare. We also find many intermediate cases that incorporate selfishness and difference aversion. Although the preferences of our subjects vary widely, the single commonest form, involving lexicographic preferences, is not explicitly recognized in previous experimental work.

Our results, which characterize behavior at the level of the individual subject, emphasize both the prominence and the heterogeneity of Pareto-damaging behaviors. Our individual-level analyses both confirm the importance of incorporating Pareto-damaging behaviors into social preferences and uncover the individual heterogeneity in these behaviors. Our paper thus contributes to a large and growing body of work on social preferences, including Loewenstein, Bazerman, and Thompson (1989), Bolton (1991), Levine (1998), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Charness and Rabin (2002), and Andreoni and Miller (2002) among others. Camerer (2003) provides a comprehensive discussion of experimental and theoretical work in economics focusing on dictator, ultimatum and trust games.

Finally, we also note that we employ a similar experimental methodology in a concurrent paper, Fisman, Kariv, and Markovits (2005). In that paper, we expand upon the experiment conducted by Andreoni and Miller (2002) who examine linear *budget sets* that vary the endowments and the prices of giving in order to test observed dictator behavior for consistency with utility maximization. Hence, while the two papers share a similar experimental methodology that allows for the collection of a rich individual-level dataset, they address very different questions.

The rest of the paper is organized as follows. Section 2 outlines some key concepts to which we refer throughout the paper and introduces the template for our analysis. Section 3 describes the experimental design and procedures. Section 4 provides the results. Section 5 discusses the results and Section 6 contains some concluding remarks. The experimental instructions are reproduced in Section 7.

2 Template for Analysis

2.1 Definitions

In this section, we define a few key concepts that we refer to throughout the paper. Assuming that the preference ordering of person $self \succsim_s$over π_s and π_o is locally nonsatiated, we say that allocation $\pi = (\pi_o, \pi_s) \in \Pi$ is *weakly Pareto superior* to allocation π' if $\pi > \pi'$, and *strictly Pareto superior* if $\pi >> \pi'$.[1] Analogous definitions apply to *weakly Pareto inferior* and *strictly Pareto inferior*. Allocation π is *weakly Pareto efficient* if there is no feasible allocation $\pi' >> \pi$, and *strictly Pareto efficient* if there is no feasible allocation $\pi' > \pi$. Figure 2 depicts the *Pareto set* in our experiment. We denote $\pi^s = (\pi^s_o, \pi^s_s)$ and $\pi^o = (\pi^o_o, \pi^o_s)$ as the $self$ and *other* strictly Pareto efficient allocations respectively.

[Figure 2 here]

When we say that an allocation is *Pareto efficient* we generally mean that it is weakly Pareto efficient. When it is possible to make one person better off and the other no worse off, we say that a *Pareto improvement* can be made. When we say *Pareto damaging* we mean not Pareto efficient. We further distinguish Pareto-damaging allocations that are weakly but not strictly Pareto efficient by defining an allocation as $self$ (*other*) *Pareto-damaging* if only $self$ (*other*) Pareto improvements can be made. More precisely:

Definition 1 Allocation π is $self$ Pareto-damaging if there exists a strictly Pareto efficient allocation π^i such that $\pi_o = \pi^i_o$ and $\pi_s < \pi^i_s$, and *other* Pareto-damaging if $\pi_o < \pi^i_o$ and $\pi_s = \pi^i_s$ for $i = o, s$.

Figure 2 also depicts the subsets of the Pareto set associated with each type of Pareto-damaging behavior. The horizontal subsets

$$\Pi^1 = \{\pi : \pi_s = \pi^s_s, 0 < \pi_o < \pi^s_o) \text{ and } \Pi^3 = \{\pi : \pi_s = \pi^o_s, \pi^s_o < \pi_o < \pi^o_o)$$

involve *other* Pareto-damaging behavior, whereas the vertical subsets

$$\Pi^2 = \{\pi : \pi_o = \pi^s_o, \pi^o_s < \pi_s < \pi^s_s) \text{ and } \Pi^4 = \{\pi : \pi_o = \pi^o_o, 0 < \pi_s < \pi^o_s)$$

involve $self$ Pareto-damaging behavior. The *competitive* allocation $\pi^c = (0, \pi^s_s)$ also involves *other* Pareto-damaging behavior. The only (weakly) Pareto efficient allocation that involves both $self$ and *other* Pareto-damaging behavior is $\pi^d = (\pi^s_o, \pi^o_s)$.

Next, we distinguish *inequality-decreasing* from *inequality-increasing* Pareto-damaging behavior.

[1] We use standard notation, so that $\pi \geq \pi'$ means $\pi_i \geq \pi'_i$ for $i = o, s$, $\pi > \pi'$ means $\pi_o = \pi'_o$ and $\pi_s > \pi'_s$ or $\pi_o > \pi_o$ and $\pi_s = \pi'_s$, and $\pi >> \pi'$ means $\pi_i > \pi'_i$ for $i = o, s$.

Definition 2 A *self* or *other* Pareto-damaging allocation π is inequality-decreasing if $d(\pi, \pi^e) < d(\pi^i, \pi^e)$ and inequality-increasing otherwise, where π^e is the *unique* equal $\pi^e_s = \pi^e_o$ Pareto efficient allocation, and π^i for some $i = o, s$ is the strictly Pareto efficient allocation that is weakly Pareto superior to π.

Hence, whether *self* or *other* Pareto-damaging behavior is inequality increasing or decreasing depends on π^e and on π^s or π^o. More precisely, whether a Pareto-damaging allocation π is inequality-decreasing or inequality-increasing depends on the difference between the distance of the allocation from the unique equal Pareto efficient allocation π^e, and the distance between π^e and the strictly Pareto efficient allocation (π^s or π^o) that is weakly Pareto superior to π. For example, if π^e is a member of Π^2 or Π^3 then all *other* Pareto-damaging allocations in Π^1 and *self* Pareto-damaging allocations in Π^4 are necessarily inequality-increasing. Finally, notice that π^d is the only allocation that is weakly Pareto inferior to both π^s and π^o. We say that π^d is inequality-decreasing if $d(\pi^d, \pi^e) < d(\pi^i, \pi^e)$ for all $i = o, s$, and an analogous definition applies to inequality-increasing.

2.2 Prototypical social preferences

Certain allocations may readily be associated with various prototypical social preferences. For ease of exposition, we use definitions that stem from the model of Charness and Rabin (2002) who consider the following simple formulation of the preferences of *self*:

$$U_s(\pi_o, \pi_s) \equiv (\rho r + \sigma s)\pi_o + (1 - \rho r - \sigma s)\pi_s,$$

where $r = 1$ ($s = 1$) if $\pi_s > \pi_o$ ($\pi_s < \pi_o$) and zero otherwise.

The parameters ρ and σ allow for a range of different social preferences:

(i) *competitive* preferences ($\sigma \leq \rho < 0$), where utility increases in the difference $\pi_s - \pi_o$, are consistent only with the competitive allocation $\pi^c = (0, \pi^s_s)$;

(ii) *narrow self-interest* or *selfish* preferences ($\sigma = \rho = 0$), where utility depends only on π_s, are consistent with any allocation π where $\pi_s = \pi^s_s$;

(iii) *difference aversion* preferences ($\sigma < 0 < \rho < 1$), where utility is increasing in π_s and decreasing in the difference $\pi_s - \pi_o$, are generally consistent with the allocations π^s and π^e if $\pi^e_s = \pi^o_s$;

(iv) *social welfare* preferences ($0 < \sigma \leq \rho \leq 1$), where utility is increasing in both π_s and π_o, are only consistent with π^s and π^o.

Notice that proportionally increasing ρ and σ indicates a decrease in self-interestedness whereas increasing the ratio ρ/σ indicates an increase in concerns for increasing aggregate payoffs rather than reducing differences in payoffs (see Charness and Rabin (2002) Appendix I). To provide a clearer intuition, Figure

3 illustrates difference aversion and social welfare preferences and depicts the range of solutions when $\pi^e \in \Pi^3$. A typical indifference curve for difference averse preferences is represented in the left panel ($MRS_{os} > 0$ for $\pi_s < \pi_o$) and for social-welfare preferences in the right panel ($MRS_{os} < 0$ for $\pi_s < \pi_o$). In these cases, the difference aversion optimum is π^s or π^e whereas the social-welfare optimum is π^s or π^o. We emphasize, however, that we do not use the model of Charness and Rabin (2002) to calibrate our experimental data, but rather to organize it and put observed behavior into perspective. Also notice that many Pareto efficient allocations are not consistent with any of the above prototypical preferences. For example, any allocation $\pi \in \Pi^3$ is not consistent with any of these preferences unless $\pi = \pi^e$.

[Figure 3 here]

Finally, person $self$ has *lexicographic* preferences for π_s over π_o if $\pi \succsim_s \pi'$ whenever $\pi_s > \pi'_s$ or $\pi_s = \pi'_s$ and $\pi_o > \pi'_o$. The lexicographic ordering is not continuous and cannot be represented by the model of Charness and Rabin (2002). Even so, it is natural and important for organizing the data. These preferences are consistent with π^s only, and thus, like social-welfare preferences (but unlike other preference prototypes), guarantee that behavior is always strictly Pareto efficient. Also note that person $self$ can, in principle, convexify the set of feasible payoffs by randomizing between π^s and π^o and evaluating outcomes according to their expected utility. There is no theoretical reason to prevent $self$ from randomizing in a way that increases aggregate payoffs and reduces differences in payoffs *ex ante*, even though resulting allocations may be very unequal *ex post*. Our experimental data, however, does not support this specification.

To summarize, our experiment allows us to distinguish among the various prototypical preferences. Most importantly, because of the non-convexity of the Pareto set, choices that reduce differences in payoffs are generically only weakly Pareto efficient. Moreover, the sharp nonlinearity of the Pareto set confronts person $self$ with an extreme relative price of giving. Hence, the step-shaped dictator game differs from the split-the-pie dictator games commonly studied in two ways. First, it does not allow for incremental efficient sacrifices that decrease inequality and it therefore provides a challenging test of $self$ and *other* Pareto-damaging difference aversion. Second, it also permits Pareto-damaging distributional preferences that increase inequality such as narrow-self interest and competitiveness. Thus, the step-shaped dictator sets "span" the set of prototypical social preferences, enabling a rigorous classification of Pareto-damaging behaviors, and providing a rigorous test of the theory.

3 Experimental Design

The experiment was conducted at the Experimental Social Science Laboratory (X-Lab) at the University of California, Berkeley under the X-Lab Master Human Subjects Protocol. The 58 subjects in the experiment were recruited from

all undergraduate classes and staff at UC Berkeley and had no previous experience in experiments of dictator, ultimatum, or trust games. After subjects read the instructions (see Section 7), the instructions were read aloud by an experimenter. No subject reported any difficulty understanding the procedures or using the computer program. Each experimental session lasted for about one and a half hours. A $5 participation fee and subsequent earnings, which averaged about $32, were paid in private at the end of the session. Throughout the experiment we ensured anonymity and effective isolation of subjects in order to minimize any interpersonal influences that could stimulate other-regarding behavior.

The procedures described below are identical to those used by Fisman, Kariv and Markovits (2005) with the exception that the set of feasible monetary payoff choices is different. Each session consisted of 50 independent decision-problems. In each decision problem, each subject was asked to allocate tokens between himself π_s and an anonymous subject π_o, where the anonymous subject was chosen at random from the group of subjects in the experiment. Each choice involved choosing a point on a graph representing a step-shaped set of possible payoff pairs.

Each decision problem started by having the computer select such a step-shaped set Π randomly from the set

$$\left\{\pi : \pi \leq \pi^s\right\} \cup \left\{\pi : \pi \leq \pi^o\right\}$$

where

$$\pi_s^o < \pi_s^s \text{ and } \pi_o^s < \pi_o^o,$$

and

$$0 \leq \pi^s, \pi^o \leq 100.$$

An example of one such set is illustrated in Figure 1 above. The sets selected for each subject in different decision problems were independent of each other and of the sets selected for any of the other subjects in their decision problems.

To choose an allocation, subjects used the mouse or the arrows on the keyboard to move the pointer on the computer screen to the desired allocation. Subjects could either left-click or press the Enter key to make their allocations. At any point, subjects could either right-click or press the Space key to find out the allocation at the pointer's current position. Notice that choices were not restricted to allocations on the frontiers so that, in theory, subjects could dispose payoffs and violate (weak) Pareto efficiency.

The π_s-axis and π_o-axis were labeled Hold and Pass respectively and scaled from 0 to 100 tokens. The resolution compatibility of the sets was 0.2 tokens; the sets were colored in light grey; and the frontiers were not emphasized. The graphical representation of the feasible sets also enabled us to avoid emphasizing any particular allocation. At the beginning of each decision round, the experimental program dialog window went blank and the entire setup reappeared. The appearance and behavior of the pointer were set to the Windows mouse default and the pointer was automatically repositioned at the origin $\pi = (0, 0)$ at the beginning of each round.

This process was repeated until all 50 rounds were completed. At the end of the experiment, payoffs were determined in the following way. The experimental program first randomly selected one decision round from each subject to carry out. That subject then received the tokens that he held in this round π_s, and the subject with whom he was matched received the tokens that he passed π_o. Thus, each subject received two groups of tokens, one based on his own decision to hold tokens and one based on the decision of another random subject to pass tokens. The computer program ensured that the same two subjects were not paired twice. Payoffs were calculated in terms of tokens and then translated at the end of the experiment into dollars at the rate of 3 tokens = 1 dollar. Subjects received their payment privately as they left the experiment.

The experiments provide us with a very rich dataset. We have observations on $58 \times 50 = 2900$ individual decisions over a variety of different step-shaped sets. Most importantly, the experimental design allows subjects to make numerous choices over a wide range of situations, and this yields a rich dataset that is well-suited to analysis at the level of the individual subject.

4 Results

4.1 Group behavior

Before examining individual subject-level decision-making, we present aggregated information on the type of allocations chosen by our subjects. To allow for small trembles resulting from the slight imprecision of subjects' handling of the mouse, all the results presented below allow for a narrow confidence interval of two tokens (i.e. for any π and $\pi' \neq \pi$, if $d(\pi, \pi') \leq 2$ then π and π' are treated as the same allocation). We generate virtually identical results allowing for a one or five token confidence interval.

We consider in turn: strictly Pareto inferior allocations; $self$ versus *other* strictly Pareto efficient allocations; $self$ versus *other* Pareto-damaging allocations; and inequality-increasing versus inequality-decreasing $self$ and *other* Pareto-damaging allocations.

We begin by examining whether subjects violated (weak) Pareto efficiency. Notice again that choices were not restricted to allocations on the Pareto frontier so subjects could dispose of payoffs by choosing allocations that are strictly Pareto inferior. Since none of the distributional preferences above would lead to Pareto violations in our experimental setup, this condition is essential to the theoretical framework that serves as a basis for our further analysis. With the narrow two token confidence interval, the data support the following result.

Result 1 *Of the 2900 allocations, only 151 allocations (5.2 percent) violate Pareto efficiency. Of these, 136 violations (90.1 percent) are concentrated in five subjects, with the remaining 14 spread among the 53 other subjects.*

Hence, although there are some differences across subjects, allocations that violate Pareto efficiency are rare. This strongly suggests that most subjects

did not have difficulties understanding the procedures or using the computer program. We omit the five subjects (ID 13, 16,17, 40, and 41) with many violations of Pareto efficiency (14, 49, 20, 23, and 30 respectively) from the analyses below. This left a total of 53 subjects (91.4 percent). We also screen out the 14 violations distributed among the remaining subjects. Thus, when we henceforth refer to total allocations, we mean the 2636 allocations after removing violations of Pareto efficiency ($2900 - 5 \times 50 - 14 = 2636$).

Among the Pareto efficient allocations, our next result provides information as to whether allocations are strictly Pareto efficient, and further distinguishes between *self* Pareto efficient allocations $\pi^s = (\pi^s_o, \pi^s_s)$ and *other* Pareto efficient allocations $\pi^o = (\pi^o_o, \pi^o_s)$. The result illustrates the relative paucity of *other* relative to *self* Pareto efficient allocations and also draws attention to the existence of many weakly Pareto efficient allocations that necessarily Pareto damage *self* or *other*.

Result 2 *Of the 2636 Pareto efficient allocations, 1874 allocations (71.1 percent) are strictly Pareto efficient. Of these, 1822 (97.2 percent) are self Pareto efficient* $\pi^s = (\pi^s_o, \pi^s_s)$ *and only 52 allocations are other Pareto efficient* $\pi^o = (\pi^o_o, \pi^o_s)$.

It should be noted that the *self* and *other* Pareto efficient allocations cannot be *Pareto ranked*. This is important in reconsidering the results of studies in which subjects choose between binary allocations that are Pareto incomparable. Compared with these typical binary dictator games, the mode at the *self* Pareto efficient allocation is much more pronounced in our study. We argue that a possible reason for the difference is that our experimental design does not force subjects into discrete and extreme choices. They may violate Pareto efficiency instead.

We next turn our attention to analyzing the frequencies of *self* and *other* Pareto efficient allocations. We use payoff calculations to measure the relative *surplus* of the strictly Pareto efficient decisions made by our subjects in the laboratory. For each decision, the relative surplus is defined by

$$(\pi^s_s - \pi^o_s)/(\pi^o_o - \pi^s_o).$$

That is, relative surplus depicts the surplus for *self* $\pi^s_s - \pi^o_s$ (i.e. the difference between the payoffs for *self* at the *self* and *other* Pareto efficient allocations) as a fraction of the surplus for *other* $\pi^o_o - \pi^s_o$. The histogram in Figure 4 shows the fraction of strictly Pareto efficient allocations that are *other* Pareto efficient as a function of log surplus. The horizontal axis measures the surplus for different deciles and the vertical axis measures the percentage of allocations in each decile that are *other* Pareto efficient. Below the bottom fifth percentile (relative surplus ≤ 0.245), of the 93 allocations, 19 allocations (20.4 percent) are *other* Pareto efficient. This declines immediately to 9 allocations out of 94 allocations (9.6 percent) between the fifth and tenth percentiles, and remains very low and roughly the same in all other deciles. Note that we divide the bottom decile in half because of the very striking decline in *other* Pareto efficient allocations within this decile.

[Figure 4 here]

Turning now to weakly Pareto efficient allocations, we similarly draw a distinction between *self* and *other* Pareto-damaging behaviors. Recall that an allocation is *self* (*other*) Pareto-damaging if only *self* (*other*) Pareto improvements can be made. As a methodological point, we emphasize again that, in contrast to response games, the dictator game rules out strategic behavior and reciprocity motivations that might trigger Pareto-damaging behaviors. Nevertheless, in the laboratory, we find that both *self* and *other* Pareto-damaging behavior occur frequently; although most of these are *other* Pareto-damaging. The next result summarizes the behavioral regularities in this regard and catalogs *self* and *other* Pareto-damaging behaviors in the laboratory.

Result 3 *Of the 2636 allocations that do not violate Pareto efficiency, 762 allocations (28.9 percent) are only weakly Pareto efficient. Of these, 213 allocations (28.0 percent) are self Pareto-damaging, all of which are in* Π^2*. The additional 513 allocations (67.3 percent) are other Pareto-damaging, of which 215 allocations (41.9 percent) are competitive* $\pi^c = (0, \pi_s^s)$*, 220 allocations (42.9 percent) are in* Π^1*, and 78 allocations (15.2 percent) are in* Π^3*. Finally, only 36 allocations (4.7 percent) are both self and other Pareto-damaging* $\pi^d = (\pi_o^s, \pi_s^o)$*.*

Figure 5 illustrates Result 3 by showing the distribution of decisions that are only weakly Pareto efficient. The horizontal axis consists of the subsets of the Pareto set and the vertical axis measures the percentages of decisions corresponding to these subsets. We present the overall distribution, as well as the distributions by three relative surplus terciles: equal surplus around 1 ($0.74 \leq (\pi_s^s - \pi_s^o)/(\pi_o^o - \pi_o^s) \leq 1.47$), high relative surplus for *self* ($(\pi_s^s - \pi_s^o)/(\pi_o^o - \pi_o^s) > 1.47$) and symmetrically high relative surplus for *other* ($(\pi_s^s - \pi_s^o)/(\pi_o^o - \pi_o^s) < 0.74$).

[Figure 5 here]

The most notable pattern in Figure 5 is the shift in the fraction of allocations in Π^2 versus Π^3 as relative surplus changes: when relative surplus for *self* is high, subjects more often choose *self* Pareto-damaging allocations $\pi \in \Pi^2$; symmetrically, allocations are more often *other* Pareto-damaging $\pi \in \Pi^3$ when relative surplus for *other* is high. It is also evident from Figure 5 that π^c allocations are infrequent, compared to Π^1 allocations, when relative surplus is close to unity.

Next, we use payoff calculations to measure the efficiency *loss* caused by the Pareto-damaging decisions, In order to assess the relative loss absorbed, the efficiency of decisions is measured separately for each subset of the Pareto set in the following way:

$$\begin{aligned} \Pi^1 &: (\pi_o - \pi_o^s)/(\pi_o^c - \pi_o^s); \\ \Pi^2 &: (\pi_s - \pi_s^s)/(\pi_s^o - \pi_s^s); \\ \Pi^3 &: (\pi_o - \pi_o^o)/(\pi_o^s - \pi_o^o). \end{aligned}$$

Thus, for each decision π the efficiency loss is defined relative to the strictly Pareto efficient allocation π^s or π^o which is weakly Pareto superior to π. For any $\pi \in \Pi^2$, for instance, the relative loss is defined in terms the payoff for $self$. More precisely, it is the difference between the actual payoff for $self$ π_s and the $self$ Pareto efficient payoff π_s^s as a fraction of the difference between the *other* Pareto efficient payoff π_s^o and the $self$ Pareto efficient payoff π_s^s. Note that strictly Pareto efficient decisions π^s and π^o have a relative loss of zero, and that π^d as well as competitive decisions π^c have a relative loss of one. Figure 6 presents the distributions of the relative loss absorbed from Pareto-damaging allocations in Π^1, Π^2 and Π^3 in the form of histograms. Comparing efficiency losses in Π^2 and Π^3, it is clear that the distribution is skewed to the right for Π^3 (i.e. allocations are more Pareto-damaging when the damage is to *other*). Thus, subjects are more willing to create greater inefficiency, to decrease inequality for example, when the cost is imposed on *other*.

[Figure 6 here]

We next turn our attention to distinguish inequality-decreasing from inequality-increasing Pareto-damaging behavior. Recall that a $self$ or *other* Pareto-damaging allocation π is inequality-decreasing (inequality-increasing) if the distance of π from the equal Pareto efficient allocation π^e is smaller (bigger) than the distance from π^e to whichever of the strictly Pareto efficient allocation (π^s or π^o) is Pareto superior to π. Since social-welfare models do not allow for Pareto-damaging behavior, and such behavior is only permitted in order to reduce inequality in difference-aversion models, our next result is particularly interesting in assessing the models' predictions.

Result 4 *All 213 $self$ Pareto-damaging allocations are inequality-decreasing. Overall, of the 513 other Pareto-damaging allocations, 424 allocations (82.6 percent) are inequality-increasing. Of these, all 215 competitive other Pareto-damaging allocations π^c and 207 out of the 220 (94.1 percent) other Pareto-damaging allocations in Π^1 are inequality-increasing. By contrast, 76 out of 78 (97.4 percent) of the other Pareto-damaging allocations in Π^3 are inequality-decreasing.*

Hence, in the laboratory, $self$ Pareto-damaging decisions always decreased inequality whereas *other* Pareto-damaging decisions more often increased inequality than decreased inequality.

4.2 Individual behavior

Our results so far tell only part of the story as they obscure individual-level heterogeneity in Pareto-damaging behaviors. To better understand the mechanisms underlying our subjects' decisions we turn now to investigating behavior at the level of the individual subject. The graphical representation of the dictator game that we employ enables us to gather richer data than has heretofore been

available and therefore makes the individual level analysis possible. Not surprisingly, particular types of allocations are concentrated in different subjects. In order to classify the preferences of individual subjects, Table 1 summarizes the distribution of Pareto efficient decisions aggregated to the subject level. The numbers measure the percentage of decisions corresponding to each subset of the Pareto set and an additional column lists the fraction of equal allocations π^e.

[Table 1 here]

Where possible, in Table 1, we also adhere to the preference classifications described in Charness and Rabin (2002) and outlined in the previous section. We find considerable heterogeneity of preferences, ranging from competitive to selfish to difference aversion to social welfare. However, the choices made by subjects with uniformly $self$ Pareto efficient allocations (i.e. $\pi = \pi^s$ in all decision-rounds) do not correspond to any of these prototypical preferences. This set of choices fits with lexicographic preference for $self$ over $other$. We also find many intermediate cases that cannot be cleanly categorized. Since the choices made by most of our subjects reflect stable underlying preferences across decision-rounds, we can report the following result.

Result 5 *Of the 53 subjects listed in Table 1, 43 of them (81.1 percent) have cleanly classifiable preferences. Of these, 26 subjects (49.0 percent) have lexicographic preferences ($\pi = \pi^s$), three subjects (5.7 percent) have competitive preferences ($\pi = \pi^c$), seven subjects (13.2 percent) exhibit selfish preferences ($\pi_s = \pi^s_s$ and $0 \leq \pi_o \leq \pi^s_o$), and seven subjects (13.2 percent) exhibit social-welfare preferences ($\pi = \pi^s$ or $\pi = \pi^o$). Of the 10 remaining subjects, nine (17.0 percent) have intermediate preferences that incorporate elements of preferences for $self$, concerns for $other$, and difference aversion. The remaining subject exhibits preferences that incorporate both difference aversion and social welfare preferences.*

The figures reported in Table 1 above provide evidence on the heterogeneity of preferences across subjects. We provide further evidence for Result 5 by separately analyzing the behaviors of each preference type, beginning with the 26 subjects whose choices correspond to lexicographic preferences. Referring to Table 1A, of these, 19 subjects choose π^s in all 50 decision-rounds, four subjects (ID 12, 28, 43, and 55) in 49 rounds, and three subjects (ID 11, 39, and 44) in 46 rounds. For the subjects that choose π^s in 46 rounds, the remaining decisions are all π^c or in Π^1. Always choosing π^s could potentially be consistent with social welfare or difference aversion preferences as well. Given the rich menu of step-shaped sets faced by each subject, however, social welfare preferences that generate $\pi = \pi^s$ for all allocations would require a great weight on $self$ (high positive ρ/σ in Charness-Rabin model). More precisely, note that the lower bound on the relative surplus $(\pi^s_s - \pi^o_s)/(\pi^o_o - \pi^s_o)$ varies by subject; it is uniformly very low and ranges empirically from 0.07 to 0.33. Accordingly, these subjects choose π^s even when π^s is relatively very inexpensive, so that if

these subjects do indeed have social welfare preferences, the weight on *other* is sufficiently low that for practical purposes, preferences may be approximated as being lexicographic for $self$ over *other*.

The allocations of these 26 subjects are also difficult to reconcile with difference aversion preferences, since this would imply choosing $\pi = \pi^e$ when $\pi^e \in \Pi^1$. Of these 26 subjects, 15 of them faced sets in which $\pi^e \in \Pi^1$, and the equal allocation π^e was never chosen. Further, while the allocations of the subjects that always choose $\pi = \pi^s$ are also consistent with perfectly selfish preferences, selfishness suggests no systematic pattern in the choice of π_o, whereas we always observe $\pi_o = \pi_o^s$. Thus, any explanation for the behavior of these subjects that relies upon social welfare, difference aversion, and selfishness seems inadequate. Overall, our data strongly suggest that choices made by these 26 subjects correspond to lexicographic preferences for $self$ over *other*, although we cannot definitely rule out social welfare preferences with extreme self-interestedness. Although lexicographic preferences incorporate self-interest and concerns about the payoffs of others, and have been employed regularly over the years by economists, previous experimental papers with which we are familiar have overlooked this useful form as a model of social preferences.

We next analyze the behavior of the seven subjects (ID 4, 25, 29, 30, 45, 56, and 58) whose choices correspond to social welfare preferences. Referring to Table 1B, of these, five subjects choose either π^s or π^o in all 50 decision-rounds, with the remaining two subjects having only a single Pareto violation. To further probe the validity of our classification, we consider whether the relative surplus $(\pi_s^s - \pi_s^o)/(\pi_o^o - \pi_o^s)$ is significantly different when subjects choose π^s relative to when they choose π^o. Table 2 summarizes the means, standard deviations, and number of observations for each of these seven subjects, according to whether π^s or π^o was chosen. For each of these subjects, relative surplus is higher when π^s is chosen, and this difference is significant at the 1 percent level in all cases. Thus, we conclude that the choices made by these subjects correspond to social welfare preferences.

[Table 2 here]

Next, we turn to the seven subjects whose choices fit with selfish preferences. Referring to Table 1C, of these, six subjects (ID 1, 3, 8, 15, 37, and 53) choose $\pi_s = \pi_s^s$ and $\pi_o \leq \pi_o^s$ in all 50 decision-rounds and the remaining subject (ID 8) chooses $\pi_s = \pi_s^s$ in 45 rounds with all of the decisions with $\pi_s < \pi_s^s$ occurring in the first 20 rounds. We say that these subjects made choices that reflect selfish preferences if the choice of π_o is random due to apparent indifference to *other*. We examined the possibility that there may be a competitive element to behaviors of these subjects by noting that if this were the case then the Charness-Rabin model predicts that π_o should increase with π_s^s, for any given π_o^s. However, a simple regression analysis indicates no relation between potential inequality and π_o for the pooled sample, and the behavior of no individual subject exhibits a significant relation between π_s^s and π_o. Table 1D summarizes the choices of the three subjects (ID 2, 49, 50) the exhibit competitive preferences.

Next, we turn to the ten subjects that exhibit $self$ Pareto-damaging behavior. Referring to Table 1E, of these, at least four subjects (ID 9, 22, 26, and 33) appear to be governed by difference aversion, as $\pi = \pi^e$ is chosen frequently. For these subjects, deviations from equality are dominated by allocations π with $\pi_s > \pi_o$, and the extent of inequality is increasing in π_s^s and decreasing in π_o^s. By contrast, inequality is uncorrelated with either π_s^o or π_o^o. Thus, these subjects made choices that may reflect a combination of selfishness and difference aversion. This is illustrated in the first column of Table 3, which reports the results of a regression predicting the extent of inequality, $d(\pi, \pi^e)$, for the $self$ Pareto-damaging allocations chosen by these four subjects.

[Table 3 here]

Additionally, five subjects (ID 24, 27, 35, 38, and 57) choose many $self$ Pareto-damaging allocations, all of which decrease inequality, though the distribution of allocations is more dominated by allocations with $\pi \neq \pi^e$. As before, the extent of inequality is increasing in π_s^s and decreasing in π_o^s. Thus, the choices made by these subjects also correspond to selfishness and difference aversion, though with a greater weight put on $self$. A linear regression analysis confirms these results. This is summarized in the second column of Table 3.

Finally, the choices remaining subject (ID 51) are distributed among π^s, π^o, Π^2, and Π^3 and thus correspond to a combination of selfish, difference aversion, and social welfare preferences. As with our social welfare subjects, the relative surplus $(\pi_s^s - \pi_s^o)/(\pi_o^o - \pi_o^s)$ is highly correlated with this subject's choice of π^s versus π^o, and as with our subjects whose choices fit with a combination of selfish and difference averse preferences, inequality in allocations $\pi \in \Pi^2$ are increasing in π_s^s and decreasing in π_o^s.

5 Discussion

In recent years, economists responding to experimental evidence have expanded the economic conception of rationality to include other-regarding preferences. Other-regarding behavior has been repeatedly and robustly demonstrated in the laboratory, and more rigorous work, including Andreoni and Miller (2002) and especially Fisman, Kariv, and Markovits (2005), suggests that other-regarding behavior is consistent with the utility maximization model and that social preferences are highly heterogeneous, ranging from utilitarian to Rawlsian to perfectly selfish.

But even though other-regarding preferences in general are increasingly well-documented and well-understood, and increasingly incorporated into economic theory, the Pareto principle continues to dominate economic conceptions of rationality. Experimental study of violations of Pareto efficiency remains in its infancy, and more theoretical models, including models that incorporate other-regarding preferences, make little room for violations of Pareto efficiency. Moreover, the lone prominent exception to this rule, difference aversion, allows only

for violations of Pareto efficiency that decrease inequality and excludes violations of Pareto efficiency that increase inequality.

Our experimental results indicate that the existing theories are inadequate. Violations of Pareto efficiency are common – nearly a third of all allocations in our experiment violate weak Pareto efficiency – and display patterns that existing models of social preferences cannot comfortably explain. Moreover, our results strongly suggest that better understanding violations of Pareto efficiency will improve explanations even of more familiar cases of social preferences.

At the aggregate level, our study reveals previously unobserved interactions among Pareto-damaging, self-interested, and inequality-averse behaviors. Unlike typical studies involving binary dictator games, in which subjects mostly choose between two Pareto-incomparable allocations, our experimental design does not force subjects into discrete and extreme choices, but allows them to violate Pareto-efficiency instead. The appeal of this option is reflected in the fact that a much higher fraction of the strictly Pareto efficient allocations are $self$ Pareto efficient in our study than in typical binary studies. Indeed, in our setting self-sacrificing allocations generally take the form of $self$ Pareto-damaging allocations. These differences between our results and the results of previous work suggest the importance of Pareto-damaging behaviors for other-regarding behavior more generally.

Moreover, our individual level analyses also challenge conventional understandings of social preferences. Although the preferences of our subjects vary widely, the single commonest form, involving lexicographic preferences, is not fully accounted for in previous experimental work. Moreover, several of our subjects display a balance of selfishness and difference aversion that leads them to make $self$ Pareto-damaging allocations that cannot be accommodated by the canonical models of social preferences encapsulated in Charness and Rabin (2002). Finally, and quite strikingly, some of our subjects, including, but not limited to, our competitive subjects, systematically display inequality-increasing Pareto-damaging behaviors. This result is of particular note in light of the fact that our experiment involves no strategic interactions and therefore eliminates reciprocity and measures purely distributive preferences.

6 Concluding Remarks

We employ a new computerized graphical representation of dictator games to study Pareto-damaging behaviors. Our experimental subjects choose allocations in step-shaped sets of possible payoff pairs that impose an extreme price of giving - so that inequality may be increased or decreased only through Pareto-damaging behaviors. This allows us systematically to classify Pareto-damaging allocations: as $self$ Pareto-damaging or *other* Pareto-damaging and as inequality-increasing or inequality-decreasing. Moreover, our experimental method enables us to collect many observations per subject, and we can therefore analyze preferences at the individual level.

The basic regularities from our experiment may be summarized as follows:

First, very few allocations Pareto damage both $self$ and $other$; that is, there are very few violations of weak Pareto efficiency. Second, between the two strictly Pareto efficient allocations, almost all maximize the payoff for $self$. Third, nearly a third of all allocations Pareto damage either $self$ or $other$ and thus are only weakly Pareto efficient. Fourth, nearly three-quarters of Pareto-damaging allocations damage $other$. Most interestingly, all $self$ Pareto-damaging allocations decrease inequality, while two-thirds of $other$ Pareto-damaging allocations increase inequality. Finally, most subjects have cleanly classifiable preferences, ranging from competitive to lexicographic to social welfare to difference aversion. We also find intermediate cases that incorporate elements of preferences for $self$, concerns for $other$, and difference aversion.

Our results emphasize both the prominence and the heterogeneity of Pareto-damaging behaviors even in a context – the dictator game – that eliminates strategic behavior and reciprocity motivations and implicates only distributive preferences. Our individual-level analyses show that the forms of Pareto-damaging behavior vary widely across subjects and, moreover, display features that are not easily accommodated by prominent models of social preferences. Our findings therefore suggest that models of social preferences must be modified in order to account for the observed choices.

Our results also suggest a number of possible extensions. In particular, a similar methodology incorporating response games could be utilized to examine the roles of strategic behavior and reciprocity in Pareto-damaging behavior. To determine which factors are important in explaining subject behavior in a variety of settings, it will be necessary to investigate a larger class of games in the laboratory. This is perhaps one of the most important topics for future research. Progress in this area requires both new theory and new experimental data. There are also many more important questions that remain to be explored using our computerized graphical representation of games. Clearly, there is much to be done and the uses of this experimental technique are far from exhausted.

7 Experimental Instructions

Introduction This is an experiment in decision-making. Research foundations have provided funds for conducting this research. Your payoffs will depend partly on your decisions and the decisions of the other participants and partly on chance. Please pay careful attention to the instructions as a considerable amount of money is at stake.

The entire experiment should be complete within an hour and a half. At the end of the experiment you will be paid privately. At this time, you will receive $5 as a participation fee (simply for showing up on time). Details of how you will make decisions and receive payments will be provided below.

During the experiment we will speak in terms of experimental tokens instead of dollars. Your payoffs will be calculated in terms of tokens and then translated at the end of the experiment into dollars at the following rate: 3 Tokens = 1 Dollar.

A decision problem In this experiment, you will participate repeatedly in 50 independent decision problems that share a common form. This section describes in detail the process that will be repeated in all decision problems and the computer program that you will use to make your decisions.

In each decision problem you will be asked to allocate tokens between yourself (Hold) and another person (Pass) who will be chosen at random from the group of participants in the experiment. The other person will not be told of your identity. Note that the person will be different in each problem. For each allocation, you and the other person will each receive tokens.

Each choice will involve choosing a point on a graph representing possible token allocations. The y-axis and x-axis are labeled Hold and Pass respectively and scaled from 0 to 100 tokens. In each choice, you may choose any Hold / Pass pair that is in the step-shaped region that is shaded in gray. Examples of regions that you might face appear in Attachment 1.

[Attachment 1 here]

Each decision problem will start by having the computer select such a step-shaped region randomly. That is, the region selected depends solely upon chance and is equally likely to be any step-shaped region. The regions selected for you in different decision problems are independent of each other and of the regions selected for any of the other participants in their decision problems.

For example, as illustrated in Attachment 2, choice A represents an allocation in which you Hold q tokens and Pass r tokens. Thus, if you choose this allocation, you will receive q tokens and the participant with whom you are matched in that round will receive r tokens. Another possible allocation is B, in which you receive s tokens, and person with whom you are matched receives t tokens.

[Attachment 2 here]

To choose an allocation, use the mouse or the arrows on the keyboard to move the pointer on the computer screen to the allocation that you desire. At any point, you may either right-click or press the Space key to find out the allocation that the pointer is at.

When you are ready to make your decision, either left-click or press the Enter key to submit your chosen allocation. After that, confirm your decision by clicking on the Submit button or pressing the Enter key. Note that you can choose only Hold / Pass combinations that are in the gray region. To move on to the next round, press the OK button.

Next, you will be asked to make an allocation in another independent decision. This process will be repeated until all the 50 rounds are completed. At the end of the last round, you will be informed the experiment has ended.

Payoffs Your payoffs are determined as follows. At the end of the experiment, the computer will randomly select one decision round from each participant to carry out. That participant will then receive the tokens that she held

in this round, and the participant with whom she was matched will receive the tokens that she passed.

Each participant will therefore receive two groups of tokens, one based on her own decision to hold tokens and one based on the decision of another random participant to pass tokens. The computer will ensure that the same two participants are not paired twice.

The round selected and your choice and your payment for the round will be recorded in the large window that appears at the center of the program dialog window. At the end of the experiment, the tokens will be converted into money. Each token will be worth 1/3 Dollars. You will receive your payment as you leave the experiment.

Rules Your participation in the experiment and any information about your payoffs will be kept strictly confidential. Your payment-receipt and participant form are the only places in which your name and social security number are recorded.

You will never be asked to reveal your identity to anyone during the course of the experiment. Neither the experimenters nor the other participants will be able to link you to any of your decisions. In order to keep your decisions private, please do not reveal your choices to any other participant.

Please do not talk with anyone during the experiment. We ask everyone to remain silent until the end of the last round. If there are no further questions, you are ready to start. An instructor will approach your desk and activate your program.

References

[1] Andreoni, J. and J. Miller (2002) "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*, *70*, pp. 737-753.

[2] Bolton, G. (1989) "A Comparative Model of Bargaining: Theory and Evidence." *American Economic Review*, *81*, pp. 1096-1136.

[3] Bolton, G. and A. Ockenfels (2000) "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review*, *90*, pp. 166-193.

[4] Camerer, C. (2003) "Behavioral Game Theory: Experiments in Strategic Interaction." Princeton University Press.

[5] Charness, G. and M. Rabin (2002) "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics*, *117*, pp. 817-869.

[6] Fehr, E. and K. Schmidt (1999) "A Theory of Fairness, Competition and Co-operation." *Quarterly Journal of Economics*, *114*, pp. 817-868.

[7] Fisman, R., S. Kariv and D. Markovits (2005) "Individual Preferences for Giving." Yale Law & Economics Research Paper No. 306.

[8] Levine, D. (1998) "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, *1*, pp. 593-622.

[9] Loewenstein, G., L. Thompson and M. Bazerman (1989) "Social utility and decision making in interpersonal contexts." *Journal of Personality and Social Psychology*, *57(3)*, pp. 426-441.

Table 1: The distribution of Pareto efficient decisions aggregated to the subject level

Table 1A: Lexicographic

ID	π^c	Π^1	π^s	Π^2	π^d	Π^3	π^o	Π^4	π^e
5	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	2.0
7	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	6.0
11	4.0	4.0	92.0	0.0	0.0	0.0	0.0	0.0	2.0
12	0.0	2.0	98.0	0.0	0.0	0.0	0.0	0.0	0.0
14	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	4.0
18	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	2.0
20	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	2.0
21	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	2.0
28	0.0	2.0	98.0	0.0	0.0	0.0	0.0	0.0	2.0
31	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
32	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	8.0
34	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
36	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	4.0
39	0.0	8.0	92.0	0.0	0.0	0.0	0.0	0.0	0.0
42	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	2.0
43	0.0	2.0	98.0	0.0	0.0	0.0	0.0	0.0	0.0
44	6.0	2.0	92.0	0.0	0.0	0.0	0.0	0.0	2.0
46	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
47	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	4.0
48	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	4.0
52	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	6.0
54	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
55	2.0	0.0	98.0	0.0	0.0	0.0	0.0	0.0	4.0
Average	0.5	0.8	98.8	0.0	0.0	0.0	0.0	0.0	2.2

Table 1B:Social welfare

ID	π^c	Π^1	π^s	Π^2	π^d	Π^3	π^o	Π^4	π^e
4	0.0	0.0	90.0	0.0	0.0	0.0	10.0	0.0	2.0
25	0.0	0.0	88.0	0.0	0.0	0.0	12.0	0.0	0.0
29	0.0	0.0	90.0	0.0	0.0	0.0	10.0	0.0	2.0
30	0.0	0.0	96.0	0.0	0.0	0.0	4.0	0.0	8.0
45	0.0	0.0	90.0	0.0	0.0	0.0	10.0	0.0	0.0
56	0.0	0.0	78.0	0.0	0.0	2.0	20.0	0.0	4.0
58	0.0	2.0	92.0	0.0	0.0	0.0	6.0	0.0	4.0
Average	0.0	0.3	89.1	0.0	0.0	0.3	10.3	0.0	2.9

Table 1C: Selfish

ID	π^c	Π^1	π^s	Π^2	π^d	Π^3	π^o	Π^4	π^e
8	24.0	6.0	60.0	6.0	0.0	4.0	0.0	0.0	8.0
19	62.0	16.0	22.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	80.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	10.0	48.0	42.0	0.0	0.0	0.0	0.0	0.0	2.0
37	0.0	87.8	12.2	0.0	0.0	0.0	0.0	0.0	0.0
53	26.5	67.3	6.1	0.0	0.0	0.0	0.0	0.0	2.0
Average	17.5	57.9	23.2	0.9	0.0	0.6	0.0	0.0	1.7

Table 1D: Competitive

ID	π^c	Π^1	π^s	Π^2	π^d	Π^3	π^o	Π^4	π^e
2	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
49	96.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0
50	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Average	98.7	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0

Table 1E: Others

ID	π^c	Π^1	π^s	Π^2	π^d	Π^3	π^o	Π^4	π^e
9	0.0	2.0	4.0	66.0	4.0	22.0	2.0	0.0	52.0
22	0.0	0.0	16.7	37.5	2.1	31.3	12.5	0.0	37.5
26	0.0	6.8	18.2	27.3	9.1	34.1	4.5	0.0	59.1
33	0.0	0.0	4.0	42.0	22.0	28.0	4.0	0.0	92.0
Average	0.0	2.2	10.7	43.2	9.3	28.8	5.8	0.0	60.1

ID	π^c	Π^1	π^s	Π^2	π^d	Π^3	π^o	Π^4	π^e
24	0.0	0.0	51.1	23.4	4.3	19.1	2.1	0.0	8.5
27	0.0	2.0	14.3	79.6	4.1	0.0	0.0	0.0	6.1
35	0.0	0.0	44.0	44.0	8.0	4.0	0.0	0.0	4.0
38	0.0	2.0	30.0	66.0	2.0	0.0	0.0	0.0	6.0
57	0.0	4.0	60.0	8.0	18.0	8.0	2.0	0.0	14.0
Average	0.0	1.6	39.9	44.2	7.3	6.2	0.8	0.0	7.7

ID	π^c	Π^1	π^s	Π^2	π^d	Π^3	π^o	Π^4	π^e
51	0.0	0.0	50.0	34.0	0.0	10.0	6.0	0.0	12.0

Table 2: The relative surplus of subjects whose choices correspond to social welfare preferences

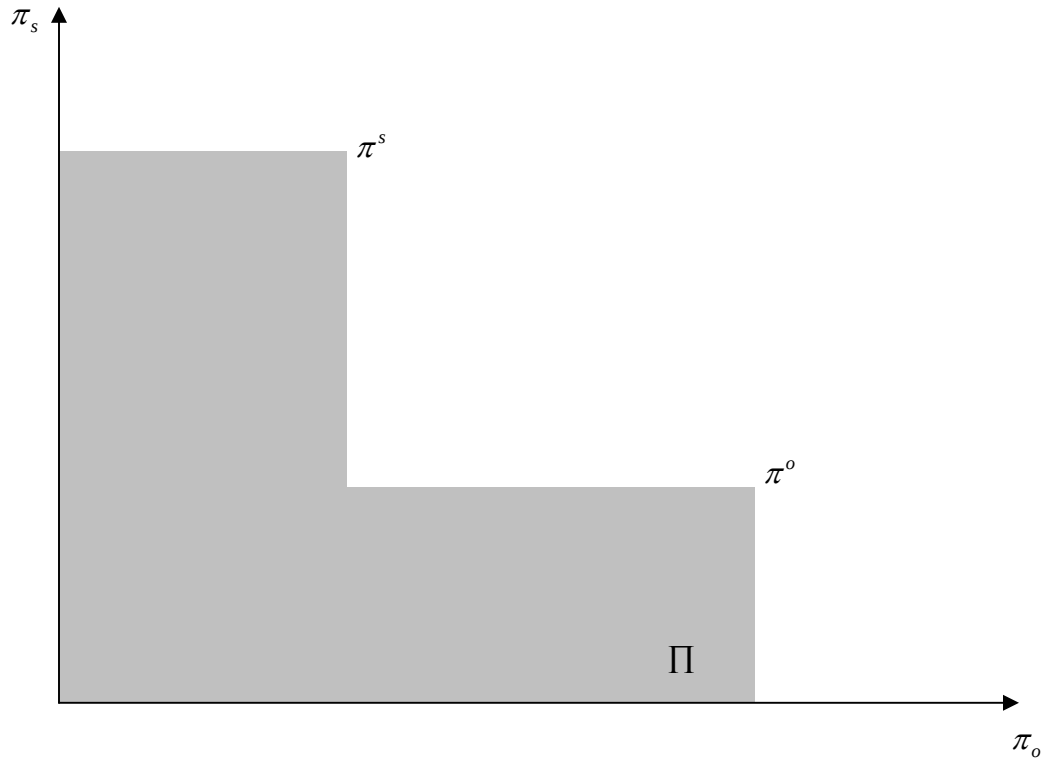
ID	π	#	Mean	SD
4	π^s	45	1.51	1.30
	π^o	5	0.36	0.37
25	π^s	44	1.51	1.04
	π^o	6	0.21	0.09
29	π^s	45	1.27	1.06
	π^o	5	0.26	0.13
30	π^s	48	1.23	0.97
	π^o	2	0.33	0.02
45	π^s	45	1.69	1.61
	π^o	5	0.33	0.24
56	π^s	39	1.84	1.02
	π^o	10	0.85	0.67
58	π^s	46	1.76	1.53
	π^o	3	0.35	0.22

Table 3: Preferences that cannot be cleanly categorized

	(1)	(2)
π^s_s	0.248***	0.481***
	(0.064)	(0.047)
π^s_o	-0.294*	-0.477***
	(0.150)	(0.090)
π^o_o	-0.041	-0.018
	(0.049)	(0.033)
π^o_s	0.099	-0.001
	(0.126)	(0.072)
# of obs.	84	109
R^2	0.48	0.69

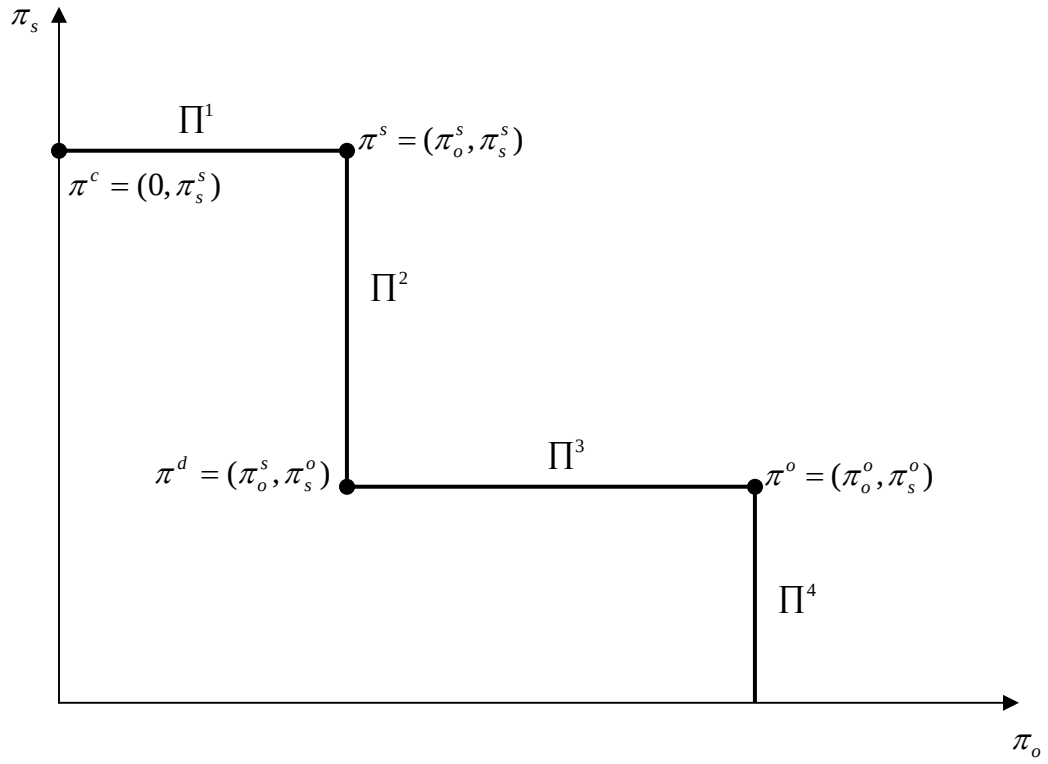
Subjects ID: (1) 9, 22, 26, 33 (2) 24, 27,35, 38, 57.
Standard errors in parentheses.
Significance level: * 10 percent, ** 5 percent, *** 1 percent

Figure 1: A step-shaped dictator set



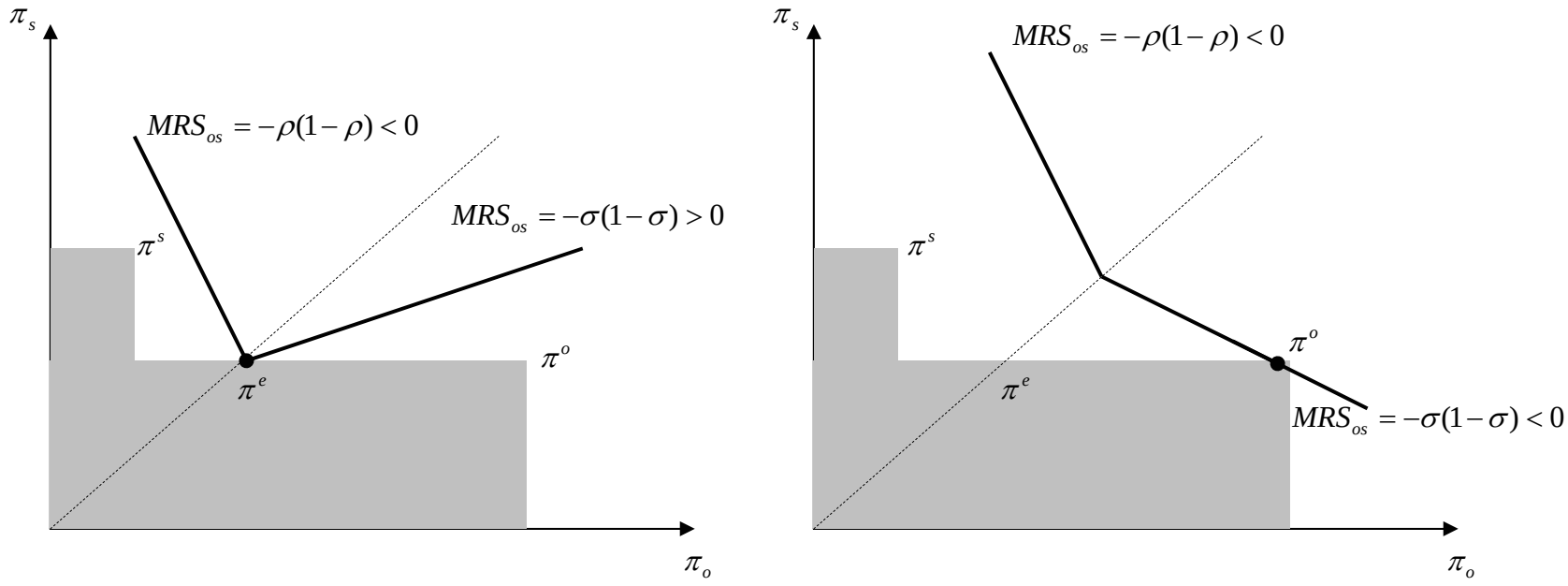
An example step-shaped set dictator set representing the feasible monetary payoff choices of person *self*. Each point $\pi = (\pi_o, \pi_s)$ corresponds to the payoffs to persons *self* and *other*, and $\pi^s = (\pi_o^s, \pi_s^s)$ and $\pi^o = (\pi_o^o, \pi_s^o)$ are the *self* and *other* strictly Pareto efficient allocations, respectively.

Figure 2: The Pareto set



The strictly Pareto efficient allocations, π^s and π^o, and the subsets of the Pareto set associated with Pareto-damaging behaviors. The horizontal subsets, Π^1 and Π^3, involve *other* Pareto-damaging behavior, whereas the vertical subsets, Π^2 and Π^4, involve *self* Pareto-damaging behavior. The allocation π^d involves both *self* and *other* Pareto demanding behaviors.

Figure 3: An example of the preferences of Charness and Rabin (2002)



Instances of social preferences and the range of solutions when $\pi^e \in \prod^3$. A typical indifference curve of a difference aversion function is represented in the left panel and of a social-welfare function in the right panel. The difference aversion optimum is π^e whereas the social-welfare optimum is π^o.

Figure 4: The distribution of log surplus

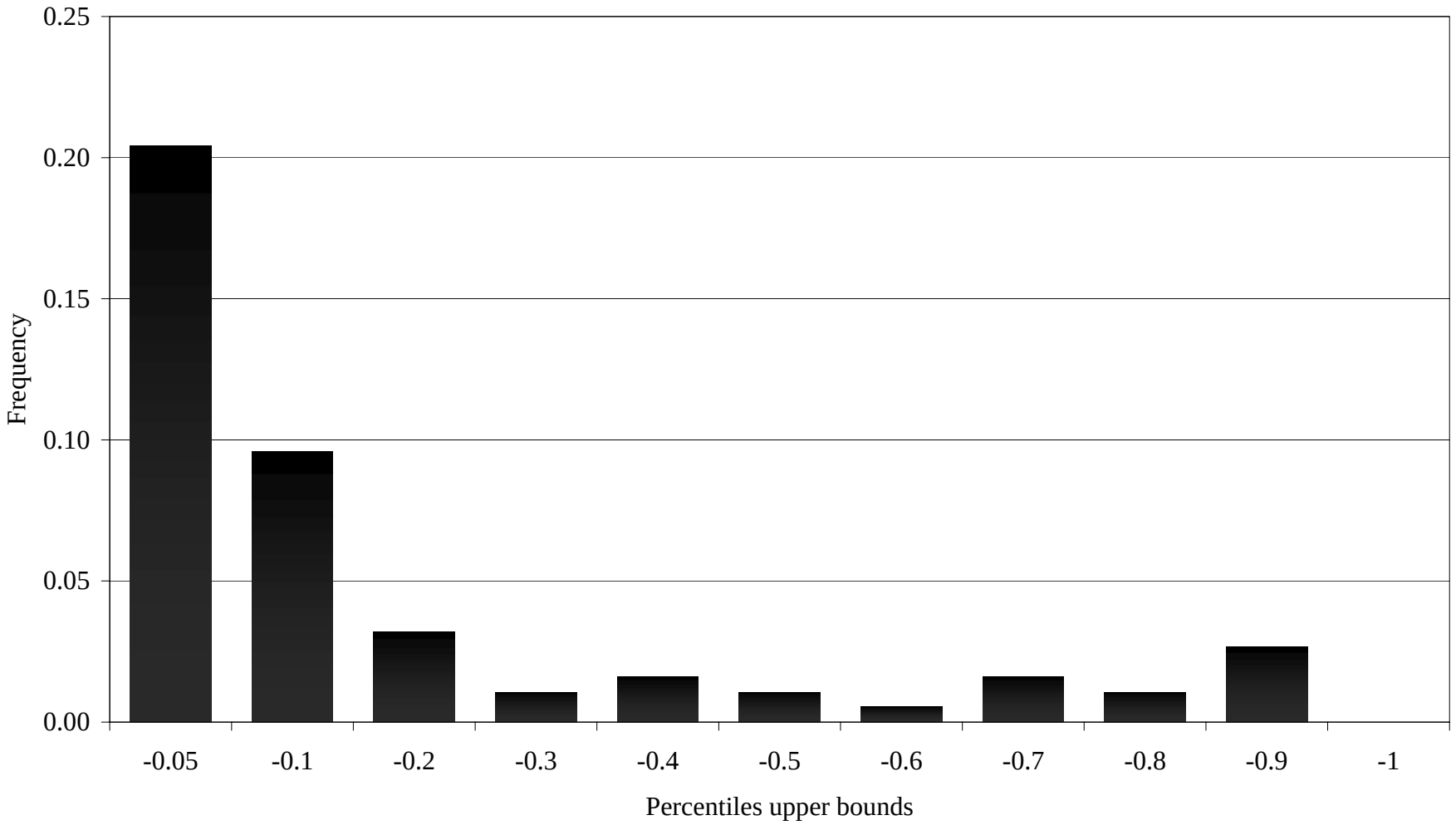
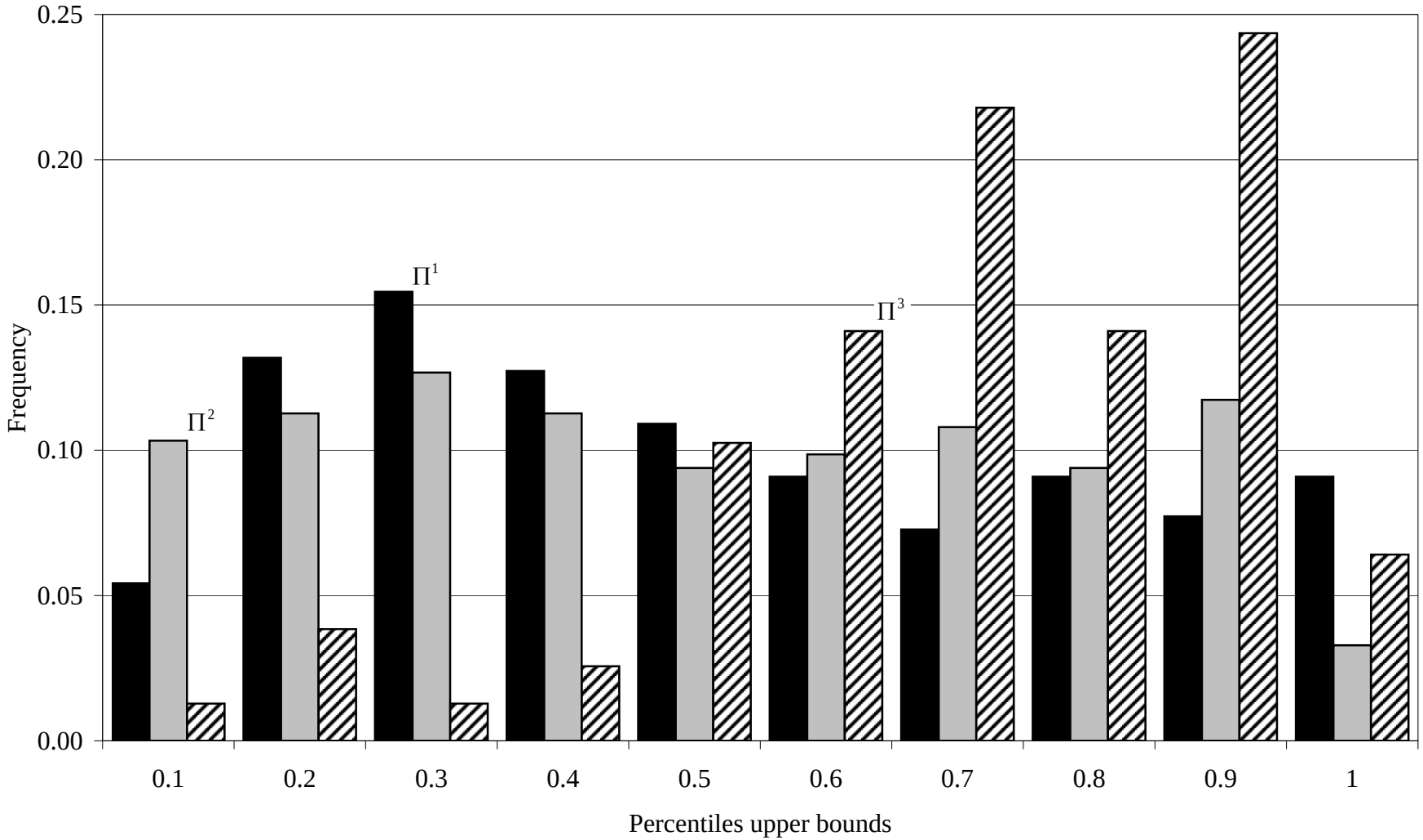
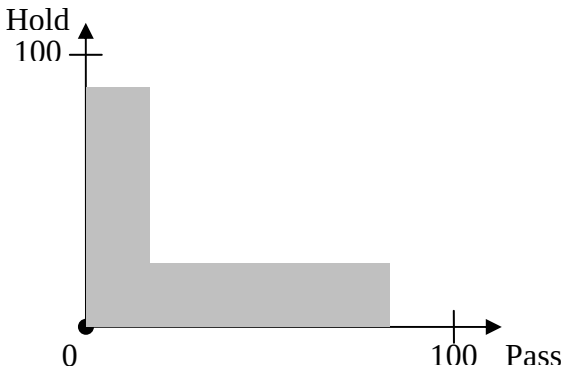


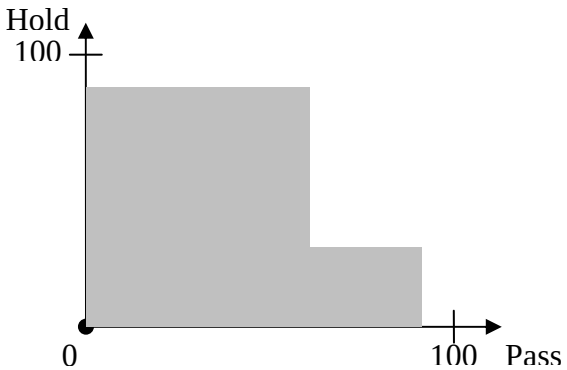
Figure 5: The distribution of weakly Pareto efficient decisions

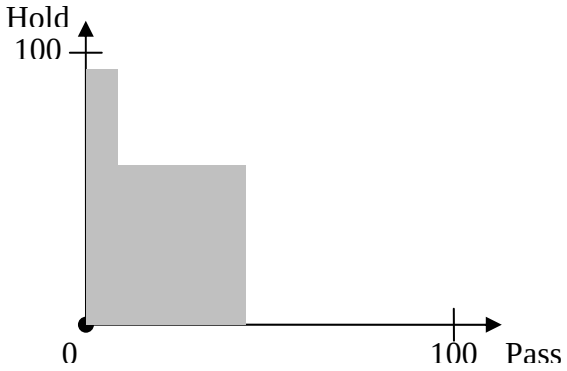
Figure 6: The distributions of the relative loss absorbed from Pareto-damaging allocations



Attachment 1







Attachment 2