Supplemental Appendix for

Social Preferences

Fundamental Characteristics and Economic Consequences

Ernst Fehr & Gary Charness

(published in the Journal of Economic Literature June 2025)

Table of Contents for Online Appendix

Appendix 1
Heterogeneity in Altruistic Distributional Preferences between Individuals and Subject Pools

This appendix describes the characterization of individual heterogeneity in terms of individuals' estimated CES utility functions. Andreoni and Miller (2002) appear to be the first who studied individual heterogeneity with the CES approach to social preferences. They recruited 176 student subjects who made between 8 – 11 choices in dictator games with varying prices of giving, allowing them to check for violations of the generalized axioms of revealed preferences (GARP). They find that less than 2% commit GARP violations, meaning that the choices of the remaining 98% can be represented by a quasi-concave utility function. They also classify individuals into one of three *predefined* categories: selfish subjects, egalitarian subjects who maximize $U_i\left(\pi_i, \pi_j\right)$ = min (π_i, π_j), and utilitarians who maximize $(0.5\ \pi_i + 0.5\ \pi_j)$. While 43 percent of their subjects display choices that perfectly fit these preference categories, the remaining 57 percent are allocated to these categories by minimizing the distance from the three pre-specified utility functions. Based on this procedure, they classify 47.2% of the 176 subjects as selfish, 30.4% as egalitarian and 22.4% as utilitarian.

To what extent do the 57% of "impure" subjects actually fit the three predefined preference categories? To answer this question, the authors estimate a representative CES function (2) for the "impure" individuals in each category. The results indicate that the estimated parameters deviate quite substantially from the parameters of the ideal types. For example, the average α' of the "impure" selfish subjects is 0.24, indicating a non-negligible deviation from selfishness, and the average ρ of the egalitarian types is –0.35 which is a long way from $-\infty$ which would indicate strict egalitarianism. While such deviations from the pure types are inevitable when people are classified into subgroups it is important to keep them in mind.

Two further observations related to Andreoni and Miller (2002) are worth mentioning. First, even those individuals who perfectly fit the selfish preference assumption in their choice data may not be perfectly selfish because the smallest relative price of giving was 0.25 – for every dollar given, the partner received $4. Thus, we do not know what would have happened if the relative price had been lower.[1] Second, 34 subjects in one of their sessions

[1] Some people may be inclined to discount situations in which the cost of altruistic acts is low, but social life is in fact pervaded by situations in which low-cost favors can be given to other people. When a colleague in the workplace asks for help, when a stranger in a city asks for directions, or when students help each other answer questions, the costs involved are often very low, while the benefits for the receiving party are high.

also faced upwards sloping budget line in (π_i, π_j)-space that involved disadvantageous inequality. Subjects could reduce inequality in these budget lines by decreasing both players' payoffs, and 8 of the 34 subjects (23.5%) actually did so. Thus, they observed some evidence in favor of inequality aversion when behind but no strong inferences can be made here given the small sample size, and the CES utility function is not capable of capturing these preferences.

The Fisman-Jakiela-Kariv-Markovits group undertook one of the most systematic characterizations of individual heterogeneity in altruistic distributional preferences in a series of papers (Fisman, Kariv and Markovits 2007; Fisman, Jakiela and Kariv 2015; Fisman et al. 2015; Li et al. 2022). Subjects in their experiments faced many different budget constraints in the material payoff space, giving them substantial power to estimate the individual preference parameters α' and ρ of the CES utility function. In Fisman, Kariv and Markovits (2007) and Fisman, Jakiela and Kariv (2015), they report the parameter estimates of 76 and 72 Berkeley undergraduates, respectively; moreover, they estimate the distributional preferences of 208 Yale Law School (YLS) students in Fisman et al. (2015) as well as of 503 US medical students in Li et al. (2017). In Figures A1a and A1b we show the cumulative distribution of the estimated weight on the self-payoff, $(1 - \alpha')$, and ρ parameter for the Berkeley and the Yale Law School students and Appendix Table A1 classifies the individuals into three categories: those close to selfishness ($1 - \alpha' > 0.95$), intermediate altruists ($0.55 \leq 1 - \alpha' \leq 0.95$) and egalitarian altruists ($0.45 < 1 - \alpha' < 0.55$). The figures and Table A1 illustrate that between 30 and 40 percent of the students put literally a weight of zero or a weight close to zero on other individuals' payoffs ($1 - \alpha' > 0.95$), while only between 8 and 25 percent of them are egalitarian altruists. Moreover, the student subject pools appear to be more oriented towards efficiency compared to equality because only between 30 and 37% of them reveal a $\rho < 0$.

Are these results from student samples generalizable to the general population? To answer this question, the Fisman-Jakiela-Kariv-Markovits group also conducted experiments with a large sample of roughly 1000 Adult Americans from the American Life Panel (ALP). The ALP subjects are broadly comparable with the US population in terms of demographic and socio-economic characteristics. To control for age, Fisman et al. (2015) use only the ALP subjects under age 40 for the comparison with the student sample. The figures show that the ALP sample under age 40 displays a much higher concern for the payoff of others (i. e., low $(1 - \alpha')$, see Figure A1a) and a much higher concern for equity compared to efficiency (i. e., a low Figure ρ, see Figure A1b) than the student sample. The following facts displayed in Table A1 are, in particular, noteworthy: (i) Among the ALP subjects under age 40, the share of individuals that are close to selfishness is only 16.2% which is much smaller than the 30-

40% among the students. (ii) The share of egalitarian altruists is with 37.2% of ALP subjects under age 40 much larger than the 8-26% among the students. (iii) The share of equality-oriented individuals ($\rho < 0$) is with 47% of ALP subjects under 40 much larger than the corresponding share among the students.[2]

These large differences between student samples and the broader population are consistent with research reported in Snowberg and Yariv (2021) and Cappelen et al. (2015). Snowberg and Yariv document that subjects from a representative sample of the US population transfer a much higher share of income (39%) to recipients in simple dictator games compared to the transfers given by a large sample of all Caltech undergraduate students, who gave only 14%. Likewise, Cappelen et al. (2015) report that in a representative sample of the Norwegian population the share transferred was 40.3% for men and 41.7% for women, while male students only gave 22.6% and female students gave 32.2%.

One noteworthy feature of the experimental design on which the data in Figure A1a and A1b and Table A1 are based is that the price of giving is randomly determined for every subject, i.e., different subjects see different prices. This means that some subjects may have seen a relatively large number of low prices for giving, which makes identification of purely selfish subjects very precise, while other subjects may have seen only a few low prices of giving, so that their assignment to the selfish versus intermediate category may be coarser.

Another important feature of the data collected by the Fisman-Jakiela-Kariv-Markovits group is that the subjects do not face upwards sloping budget lines in (π_a, π_b)-space. Thus, by construction, the subjects do not face a situation in which they can decrease both players' payoffs to reduce disadvantageous inequality. Given this restriction, the CES approach is a powerful tool for identifying *altruistic* distributional preferences, but it cannot capture spiteful, envious, or inequality averse preferences[3].

[2] The FJKM group also shows that the much higher degree of other-regardingness and the much higher equality orientation of the broad population sample does not depend on socio-economic status. In other words, the individuals with high education and income in the ALP sample (N = 152) display very similar parameters compared to the rest of the ALP sample.

[3] In Fisman, Kariv, Markovits (2007), the authors had budget constraints with vertical and horizontal segments, but their student subjects never made pareto-damaging choices on these segments, which led the authors to believe that inequality aversion is not important.

Figure A1a

The estimated weight on self-payoff (1 – α') among students and in a broad sample of the US population under age 40. High (1 – α') means a low concern for others' payoff.
(based on data from FKM 2007, FJK 2015, FJKM 2015)[4]

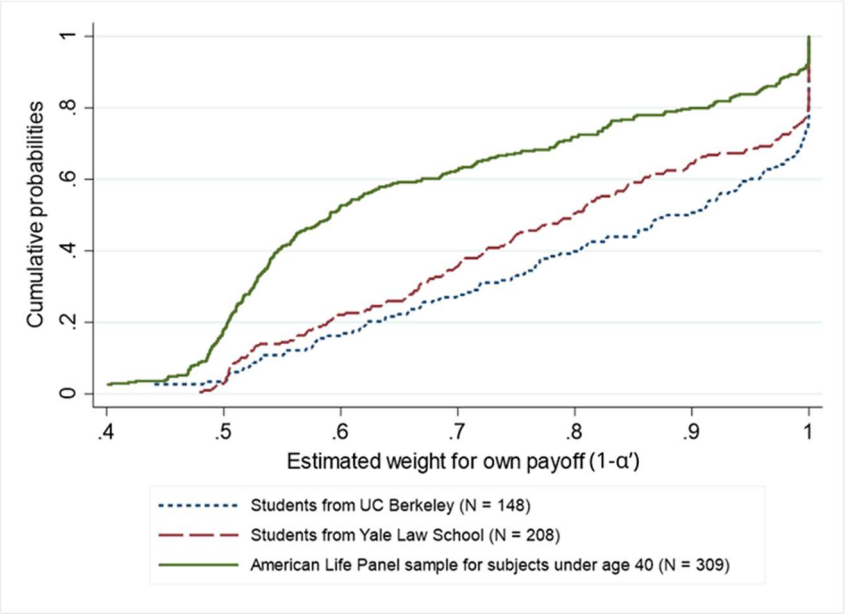
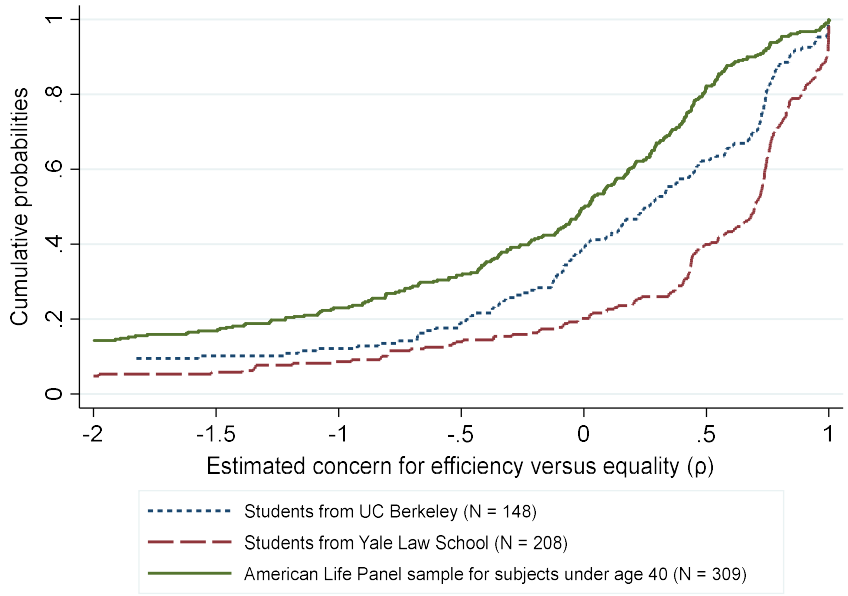


Figure A1b

The estimated weight of efficiency relative to equality concerns (ρ) among students and a broad sample of the US population under age 40. High ρ means a low concern for equality.
(based on data from FKM 2007, FJK 2015, FJKM 2015)



[4] FKM (2007) indicates Fisman, Kariv and Markovits (2007), FJK (2015) indicates Fisman, Jakiela and Kariv (2015) and FJKM indicates Fisman, Jakiela, Kariv and Markovits (2015). See also Table A1 for the type classification that follows from the estimates displayed in Figures 2a and 2b.

Table A1: Empirical Properties of Altruistic Distributional Preferences

Study	**Subject Pool**	**Egalitarian altruism $0.45 < 1-\alpha' < 0.55$**	**Intermediate altruism $0.55 \leq 1-\alpha' \leq 0.95$**	**Close to Selfishness $1-\alpha' > 0.95$**	**$\rho < 0$**
FKM 2007 & FJK 2015	N = 148 UC Berkeley students	8.1%	49.3%	39.9%	37.%
FJKM 2015	N = 208 Yale Law School Students	14.5%	53.9%	31.8%	20.3%
JDK 2017	N = 503 Students from US medical schools	25.7%	41.5%	28.2%	29.2%
FJKM 2015	N = 309 Adult Americans under 40 (ALP subjects)	37.2%	42.7%	16.2%	47.3%
	N = 693 Adult Americans over 40 (ALP subjects)	27.7%	50.5%	16.0%	57.0%
LDK 2017	N = 208 US Physicians	36.8%	42.8%	15.1%	48.3%

Note. The table shows key components of the distribution of individuals' estimated weights (α') on other persons' payoffs based in studies co-authored by D (Dow), F (Fisman), J (Jakiela), K (Kariv), L (LI) and M (Markovits). Thus. FKM (2007) indicates, e. g., the paper by Fisman, Kariv and Markovits (2007). The estimates are based on the assumption that distributional preferences can be captured by a CES utility function like in equation (5) of the paper. Each experimental subject made distributional choices for 50 randomly chosen budget sets. The efficient frontier of the budget set (i.e., the "budget line") was always negatively sloped such that one cannot measure the willingness to pay *to reduce* others' income for the sake of equality ("inequality aversion"). However, the CES function enables the identification of individuals' preference for equality *within the class of altruistic preferences* with the parameters α' and ρ. $\alpha' = 1/2$ indicates that individuals put equal weight on their own and the other players' payoff, and $\rho < 0$ implies that the income share spent on others' payoff rises as the price of giving rises, i.e., subjects are equality-oriented ($\rho < 0$) and not efficiency-oriented ($0 < \rho < 1$).

Appendix 2

The Equality Equivalence Test

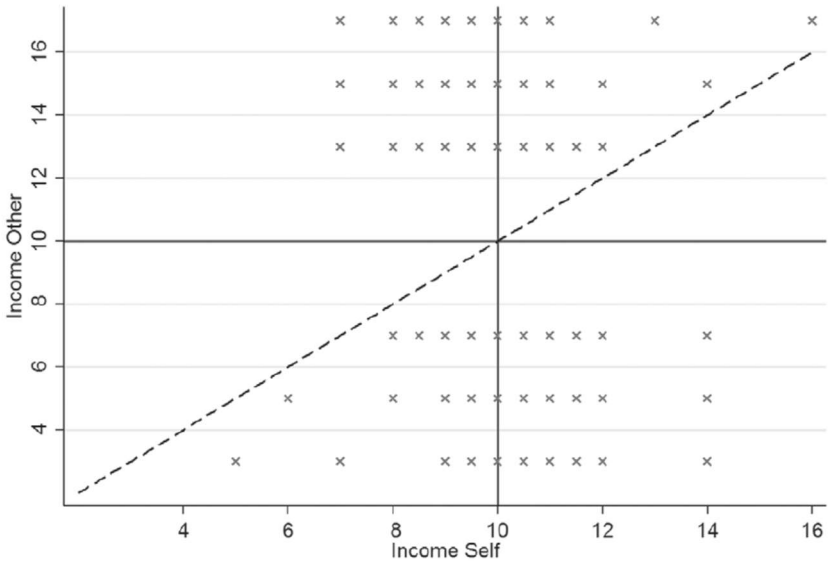
In the equality equivalence test subjects are presented choice lists in the domain of advantageous payyofs (A-lists) and the domain of disadvantageous payoffs (DA-lists). In a disadvantageous list (DA-list, see Figure A2 below), the equal payoff allocation E is always paired with a list of alternative allocations in which the other subject's payoff is kept constant at a level of $\pi_j > \pi_i$, while π_i systematically varies across alternative allocations. In an advantageous list (see Figure A2), E is always paired with a list of alternative allocations in which the other subject's payoff is kept constant at a level of $\pi_j < \pi_i$ while π_i systematically varies across alternative allocations.

Starting the binary choice list with the choice between the (π_i, π_j)-combination and E where π_i is *lowest* (and hence, below the egalitarian payoff, see Figure A1), the decision maker is more benevolent towards the other subject (i.e., willing to pay to increase the other's payoff) in the DA domain, the earlier he or she moves from E towards an alternative allocation (π_i, π_j). In the advantageous domain, the decision maker is more benevolent if he or she, starting the binary choice list with the choice between the (π_i, π_j)-combination and E where π_i is *highest* (and hence, above the egalitarian payoff), moves earlier to the equal payoff allocation E. However, the EET can also identify inequality aversion in the DA domain because some binary choice pairs essentially imply a choice on a positively sloped "budget line". Likewise, the EET can also identify positively sloped indifference curves in the A domain ("spite") because some binary choice pairs in this domain are located on positively sloped "budget lines".

A potential drawback of the EET is that the equal payoff allocation is part of every binary choice the subjects face, which may render equality very salient and thus induce a behavioral bias towards equality. However, a study by Krawczyk and Lee (2021) indicates that the results are robust to the introduction of a reference allocation that does not involve equality. In addition, the results of the EET by Kerschbamer (2015) indicates that 48.9% of his student subjects reveal selfish preferences (see Table 1 below), which is even higher than the 39.9% of selfish students in Fisman, Kariv and Markovits (2007) or the 31.8% of selfish students in Fisman et al. (2015). Likewise, Table 1 presents the data from several other studies with student samples that indicate a relatively high share of selfish subjects that approaches 60% in some student samples. Moreover, among the student subjects with other-regarding distributional preferences, those with altruistic preferences are far more prevalent

compared to inequality averse or envious preferences. The share of altruistic student subjects varies between 28 and 48%, while the share of inequality averse subjects is between 7 and 12%. Typically, envious/spiteful subjects are the least frequent across the student data with 3-10%.

Figure A2: Choice Alternatives in the Equality-Equivalence Test



The figure illustrates how the Equality Equivalence Test (EET) works by depicting the alternatives to the equal payoff allocation which is at (10, 10). The figure is taken from Kerschbamer and Müller (2021). It shows three binary choice lists in the disadvantageous domain (DA-lists) and three lists in the advantageous domain (A-lists). The DA-lists enable the identification of the slope of a subject's indifference curve in the DA domain (α), while the A-lists enable identification of the slope in the A domain (β).

Appendix 3: Distributional Preferences under Risk

When individuals care for others' payoffs, a whole new set up of questions arises if outcomes are risky. A key issue concerns the question whether people care for others' expected payoffs or for their realized payoffs. This also concerns the issue whether individuals care for equality of opportunity, i.e., have a preference for lower inequality in ex-ante expected payoffs or whether they have a preference for more equal ex-post realized payoffs. Another issue when risk is present is how individuals' own risk preferences and their beliefs about others' risk preferences affect their other-regarding behavior.

The problem of ex-ante expected payoffs versus ex-post realized payoffs comes into sharp focus in a dictator game that involves the sharing of chances to win an indivisible resource that has a value of R = 100 for both parties. The dictator chooses *x,* which determines the probability $\frac{x}{100}$ with which the recipient wins R, while the dictator wins R with probability $(1 - \frac{x}{100})$. Here, equality of opportunity implies the equalization of chances but there will always be inequality ex-post. An individual with utility function $U(\pi_a, \pi_b)$ that obeys the plausible restriction $U(R, 0) > U(0, R)$ will always choose $x = 0$. Not only inequality averse players, but players with Charness-Rabin preferences as well, may plausibly obey the restriction $U(R, 0) > U(0, R)$ and thus choose $x = 0$.

This prediction contrasts, however, with the results of experiments showing that many dictators are willing to transfer some chance of winning to the recipients (Krawczyk and Le Lec 2010; Brock, Lange and Ozbay 2013). Models that are solely based on the realized ex-post payoffs have a hard time explaining this fact, whereas models in which players also care about the ex-ante expected payoffs of others can explain it.

Now suppose that the above-described game is slightly changed so that the payoff to the two players is no longer exclusive, i.e., if the dictator transfers a chance *x*, then the dictator wins R with probability $(1 - \frac{x}{100})$ and the recipient can simultaneously also win R with $\frac{x}{100}$, i.e., there are two independent draws. Note that there may not be any ex-post inequality in this game because both players can end up with 0 or with R. Therefore, inequality averse dictators have less reason to worry about inequality, implying that they are more likely to be willing to share chances with the recipient. Krawcyk and Le Lec (2010) indeed show that dictators transfer more chances in the dictator game with independent draws compared to the game with exclusive payoffs. This result suggests that players also care about ex-post payoffs.

Further evidence for the relevance of ex-post payoffs is provided by Brock, Lange and Ozbay (2013), who designed six different dictator games where they systematically varied the risk for the dictators and the recipients across games in such a way that if players' cared only about ex-ante expected payoffs, they would behave identically across all six games. Their find treatment differences, however, that are indicative for the relevance of ex-post payoff concerns. For example, subjects in the standard dictator game without any risk (and a dictator endowment of 100) transfer a significantly higher x to the recipient compared to a dictator game where the dictator's payoff is still certain, but a transfer of x gives the recipient a payoff of 100 with probability $\frac{x}{100}$. Note that a positive transfer x in the game where the recipient faces a risky payoff implies that the dictator may end up with a lower payoff than the recipient. Inequality averse dictators, who care for ex-post inequality, will thus tend to give less in the risky dictator game.[5] Another key result documented in Brock, Lange and Ozbay (2013) is that subjects' giving in the standard dictator game is highly predictive for their willingness to equalize ex-ante expected values in dictator games involving risks.

The question whether subjects care for equality of opportunity or for equality of ex-post payoffs was also addressed in Cappelen et al. (2013). In their experiments, there was first a risk-taking phase and then a distribution phase. Subjects made 4 decisions in the risk-taking phase between the payoff y of a sure alternative ($y \in \{25, 200, 300, 400\}$ and a 50:50 chance of receiving nothing or 800 NOK. In the distribution phase, each subject was paired sequentially with 8 different subjects who participated in the risk-taking phase, and one of the four risk-taking problems was drawn randomly for each pair. Then, an "impartial" spectator, who was informed about subjects' choices and outcomes in the drawn risk-taking problem, was asked to distribute the pair's total earnings between the two subjects.

Before presenting the results, it is important to emphasize that complete equality of opportunity existed between the two paired subjects in the risk-taking phase. If spectators redistribute ex-post from the richer to the poorer subject, they thus explicitly express a preference for less ex-post inequality. Almost all of the spectators' redistributive choices

[5] Alternatively, because the certainty equivalent of a given transfer x is less valuable for risk averse recipients, dictators who care for the total payoff may give less in the risky dictator game. However, based on this logic risk averse dictators should give *more* in a dictator game in which their own payoff is risky – they receive a payoff of 100 with probability $\left(1 - \frac{x}{100}\right)$ – while the recipient receives the transfer x with certainty. The reason is that a transfer of x decreases the certainty equivalent of the dictator's payoff by less than x, i.e., giving is surplus-enhancing. The evidence strongly suggests the opposite, as dictators give much less in this game compared to the standard dictator game (Freundt and Lange 2017). Moreover, Freundt and Lange also find that the dictators who believe that recipients are risk averse do not give less to the recipients.

involved redistribution from the poorer to the richer subject.[6] If the pair consisted of two risk takers where one was lucky while the other was unlucky, the spectators strongly redistributed from the lucky to the unlucky one – they chose the equal split in more than 40% of the cases and they did not redistribute at all in only roughly 30% of the cases. In contrast, if an unlucky risk-taker was paired with an individual who chose the safe option, the unlucky risk-taker received much fewer transfers and the equal split was only chosen in roughly 15% of the cases. Spectators thus made the unlucky risk takers more responsible for their choices compared to a situation where both were unlucky. Finally, there is also a substantial amount of redistribution when a lucky risk-taker is paired with an individual who chose the safe payoff, but the lucky risk-taker was nevertheless given a higher payoff in roughly 80% of the cases.

Thus, taken together, the literature suggests that subjects on average care about both ex-ante equality of opportunity and ex-post equality of outcomes but there is strong heterogeneity in the weight that individual subjects put on the different conceptions of equality. Cappelen et al. (2013) estimate a mixture model that enables them to assign individuals to three different types – individuals who care only for ex-post equality ("ex-post egalitarians", EPs), individuals who do not care about ex-post equality ("ex-ante egalitarians", EAs), and individuals who care about ex-post equality among those who made the same choice in the risk-taking task ("choice egalitarians", CEs). Roughly 30% of their subjects (students from the Norwegian School of Economics) are EPs, 27% are CEs and 43% are EAs.

In this section we have so far mainly dealt with the question how social preferences are affected by outcome risks. However, the perceived sources of inequality may also be subject to risk and uncertainty. If individuals do not know whether a particular inequality is due to luck or differential performance, how does this affect their willingness to redistribute income? Cappelen et al (2022) study this situation, and document that this kind of uncertainty can push meritocrats towards behaving more egalitarian – with more risk averse spectators exhibiting a stronger drive towards egalitarian behavior.

Finally, we deal with the question how to combine concerns for equality of opportunity and equality of outcomes in theoretical modelling. Saito (2013) addresses this issue, providing an axiomatic foundation for "expected inequality-averse" preferences. Individuals

[6] In case that a lucky risk-taker met a subject who chose the safe payoff it would have been possible to redistribute from the poorer to the richer subject.

with such preferences put a weight of δ ($0 \leq \delta \leq 1$) on preferences for equality of opportunity and a weight (1-δ) on preferences for equal ex-post outcomes.[7]

To make things concrete, let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ denote an allocation of material payoffs to individuals. Assume that there are *m* different states of the world, each one of which is obtained with probability p_s, $s \in \{1, \ldots, m\}$, and denote the allocation obtained in state s by $\boldsymbol{x^s} = (x_1^s, x_2^s, \ldots, x_n^s)$, then the *expected* material payoff allocation is given by

$$E(\boldsymbol{x}) = \sum_{s=1}^{m} p_s \boldsymbol{x^s} = (\sum_{s=1}^{m} p_s x_1^s, \sum_{s=1}^{m} p_s x_2^s, \ldots, \sum_{s=1}^{m} p_s x_n^s),$$

where $\sum_{s=1}^{m} p_s x_i^s$ denotes the expected material payoff of individual *i* across states. Likewise, the allocation of expected utilities is given by

$$E(U(\boldsymbol{x})) = \sum_{s=1}^{m} p_s U(\boldsymbol{x^s})$$

Saito shows that if and only if a decision-maker obeys "his" axioms, the preferences of a decision-maker are represented by the following preference function V:

$$V = \delta U(E(\boldsymbol{x})) + (1-\delta) E(U(\boldsymbol{x}), \qquad (11)$$

where $U(\boldsymbol{x})$ is given by the Fehr-Schmidt Utility function. Thus, the utility of an expected inequality averse player is affected by the inequalities in the expected material payoffs with weight δ and by the inequalities in realized ex-post payoffs with weight $(1-\delta)$. It is also noteworthy that the preference function (8) also applies under further plausible assumptions if $U(\boldsymbol{x})$ is given by Charness-Rabin type preferences.

It is easy to see that an individual who puts a sufficiently high weight δ on equality of opportunity is willing to share the chances of receiving an indivisible resource in a dictator game although this creates chances for high ex-post inequality. Overall, however, the Saito model has undergone very little empirical testing. For example, it would be interesting to know to what extent the behavior of individual subjects in the six different treatment conditions of Brocks, Lange and Ozbay (2013) are consistent with the Saito model and which parameters *(α, β, δ)* explain their behaviors.[8] To our knowledge, there is no paper that jointly estimated δ and the parameters in $U(\boldsymbol{x})$. One complication in applying (8) to data is that the distributional preference models – such as Fehr-Schmidt or Charness-Rabin – assume risk neutrality, but it is well known that risk aversion also exists at the typical experimental stake

[7] Several other authors have also provided axiomatic foundations of inequality averse preferences (Neilson 2006; Rohde 2010) but none of them involves preferences for equality of opportunity.

[8] Recall that in their experiments an individual with $\delta = 1$ would behave identically across all treatments. Thus, behavioral variation across treatments may provide at least some qualitative insights with regard to the parameter constellations that may explain their data.

levels. This means that behavior in distributional problems under risk is affected by a complicated mix of risk aversion as well as by preferences for equality of opportunity and other-regarding preferences for ex-post outcomes.[9]

[9] Cettolin, Riedl and Tran (2017) and Freundt and Lange (2017) have independent measures of dictators' and recipients risk aversion and can relate them to the dictators' behavior in risk-involving dictator games. Cettolin, Riedl and Tran (2017) show that dictators' risk aversion strongly predicts lower transfers in both dictator games that render the payoff of the recipients risky and in dictator games that render the payoff of the dictators risky. Freundt and Lange (2017) also show that a rise in dictators' risk aversion is associated with a decline in generosity in games where the dictators' payoff is subject to risk.

Appendix 4: Payoff Matrices used in Bolton, Brandts and Ockenfels (1998)

Table 1. Payoff rows for the three test matrices (payoffs in Spanish pesetas).

	*c*1	*c*2	*c*3	*c*4	*c*5	*c*6
t	*C* gets 2050	*C* gets 2000	*C* gets 1950	*C* gets 1900	*C* gets 1850	*C* gets 1800
	R gets 800	*R* gets 1000	*R* gets 1200	*R* gets 1400	*R* gets 1600	*R* gets 1800
m	*C* gets 1650	*C* gets 1600	*C* gets 1550	*C* gets 1500	*C* gets 1450	*C* gets 1400
	R gets 900	*R* gets 1100	*R* gets 1300	*R* gets 1500	*R* gets 1700	*R* gets 1900
b	*C* gets 1250	*C* gets 1200	*C* gets 1150	*C* gets 1100	*C* gets 1050	*C* gets 1000
	R gets 1000	*R* gets 1200	*R* gets 1400	*R* gets 1600	*R* gets 1800	*R* gets 2000

Appendix 5 – The Stability of Social Preferences

In this appendix, we review evidence that examines the extent to which social preferences are relatively stable. Measuring the stability of social preferences over time appears straightforward as long as the measurement tools indeed deliver a preference measure and not merely a behavioral measure that is confounded by beliefs and other types of preferences (as discussed in the section on external validity), and as long as the measurement tool at different points in time is identical. In addition, the preference measure is ideally not just based on a single behavioral measure like the choice of the transfer in a standard dictator game but instead on many choice situations across which the costs and benefits of the transfer vary. Otherwise, the recovered preferences contain a lot of measurement errors and noise, which may generate spurious preference instability.

Measuring social preferences across contexts is trickier because the notion of stability is theory-dependent. To illustrate this point, consider the behavior of responders in two versions of the ultimatum game (Blount 1995). In version 1, a random mechanism determines the first-mover's offer exogenously while the first-mover herself makes the offer in version 2. Suppose that the responders are negatively reciprocal but *not* inequality averse. Then responders reject low offers in version 2 of the game but not in version 1 because a low offer does not indicate an unkind intention in version 1 but it does so in version 2 of the game. If one erroneously assumes that responders are inequality averse, one would conclude that the responders' inequality averse preferences are highly unstable because inequality averse responders should reject low offers regardless of whether they are randomly determined or volitionally chosen. However, if one correctly assumes that the responders are negatively reciprocal, their *change in behavior* across the two games is exactly what a stable preference for negative reciprocity predicts. Thus, the extent to which one can interpret changes in behavior across different contexts as changes in preferences is strongly dependent on the assumption about the underlying psychological mechanism. For this reason, care needs to be exercised when preference stability is assessed by examining behaviors across contexts.

With the above caveats in mind, what does the evidence on the stability of social preferences show? Bruhin et al. (2019) estimated the structural parameters twice for advantageous (β') and disadvantageous (α') inequality aversion in a sample of N = 196 students three months apart with the same experimental paradigm. They found that the intertemporal correlation of individuals' α' is 0.48 while the correlation for β' is 0.56.

Fehr®Epper®Senn (2022) also measured individuals' social preferences in a broad Swiss sample (N = 415) at two points in time that were three years apart (in 2017 and 2020).

The subjects faced the exact same large set of budget lines which makes it possible to study preference stability (i) at the level of choice for individual budget lines, (ii) at the level of individuals' estimated structural preference parameters and (iii) at the level of individuals' assignments to different preference types. At the choice level, roughly 55% of the choices are perfectly identical across time points and 67% of the choices are identical or coincide with the closest neighboring allocation on the budget line. At the level of individuals' structural parameters, they find an intertemporal rank correlation of 0.458 for α' and 0.428 for β'. Finally, at the level of type assignment, they find that 68% of the individuals are assigned to the same preference type (altruistic, inequality averse, selfish) across the two points in time, and that among the individuals classified as other-regarding (altruistic or inequality averse) in 2017, 89% are again classified as other-regarding in 2020. Among the individuals classified as selfish in 2017, 60% are again classified as selfish in 2020.

Moreover, two waves of the German Internet Panel implemented the same equality equivalence Test (Kerschbamer and Muller 2020). In total N = 2583 individuals participated twice in this test, 2 years apart (2016 and 2018). This permits an analysis of the stability of individuals' assignment to four pre-defined preference types (selfish, altruistic, inequality averse, envious; see Table 2). This analysis shows that 60% of individuals remain assigned to the same preference type across the two years, and that among the 76% of individuals who were classified as altruistic or inequality averse in 2016, 84.5% were again assigned to these two preference types.

Chuang and Schechter (2015) also report significantly positive intertemporal correlations between 0.21 and 0.32 involving survey measures of negative reciprocity taken in 2007, 2009 and 2010. Likewise, Carlsson, Johansson-Stenman and Nam report significantly positive intertemporal correlations of social preference related behaviors (voluntary money and labor contributions to a natural public good) at four different points in time spread across six years.

Thus, taken together, the data suggest a reasonable degree of stability in social preference when measured at the level of choices, structural parameters, or preference type assignment. However, the data also suggests a non-negligible degree of noisiness and/or measurement error. Nevertheless, the observed degree of stability appears sufficiently strong to suggest that workers with different degrees of prosociality may self-select into different sectors or to make it worthwhile for employers to screen potential employees based on certain social preference characteristics.

References

Brock, J. M., A. Lange, and E. Y. Ozbay. "Dictating the Risk: Experimental Evidence on Giving in Risky Environments." *American Economic Review* 103, no. 1 (2013): 415-37.

Bruhin, A., E. Fehr, and D. Schunk. "The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences." *Journal of the European Economic Association* 17, no. 4 (2019): 1025-69.

Cappelen, A. W., J. Konow, E. O. Sorensen, and B. Tungodden. "Just Luck: An Experimental Study of Risk-Taking and Fairness." *American Economic Review* 103, no. 4 (2013): 1398-413.

Cappelen, A. W., J. Mollerstrom, B. A. Reme, and B. Tungodden. "A Meritocratic Origin of Egalitarian Behaviour." *Economic Journal* 132, no. 646 (2022): 2101-17.

Cappelen, A. W., K. Nygaard, E. O. Sorensen, and B. Tungodden. "Social Preferences in the Lab: A Comparison of Students and a Representative Population." *Scandinavian Journal of Economics* 117, no. 4 (2015): 1306-26.

Cettolin, E., A. Riedl, and G. Tran. "Giving in the Face of Risk." *Journal of Risk and Uncertainty* 55, no. 2-3 (2017): 95-118.

Chuang, Y. T., and L. Schechter. "Stability of Experimental and Survey Measures of Risk, Time, and Social Preferences: A Review and Some New Results." *Journal of Development Economics* 117 (2015): 151-70.

Fisman, R., P. Jakiela, and S. Kariv. "How Did Distributional Preferences Change During the Great Recession?" *Journal of Public Economics* 128 (2015): 84-95.

Fisman, R., P. Jakiela, S. Kariv, and D. Markovits. "The Distributional Preferences of an Elite." *Science* 349, no. 6254 (2015).

Fisman, R., S. Kariv, and D. Markovits. "Individual Preferences for Giving." *American Economic Review* 97, no. 5 (2007): 1858-76.

Freundt, J., and A. Lange. "On the Determinants of Giving under Risk." *Journal of Economic Behavior & Organization* 142 (2017): 24-31.

Kerschbamer, R., and D. Muller. "Social Preferences and Political Attitudes: An Online Experiment on a Large Heterogeneous Sample." *Journal of Public Economics* 182 (2020).

Krawczyk, M., and F. Le Lec. "'Give Me a Chance!' An Experiment in Social Decision under Risk." *Experimental Economics* 13, no. 4 (2010): 500-11.

———. "How to Elicit Distributional Preferences: A Stress -Test of the Equality Equivalence Test." *Journal of Economic Behavior & Organization* 182 (2021): 13-28.

Li, J., L.P. Casalino, R. Fisman, S. Kariv, and D. Markovits. "Experimental Evidence of Physician Social Preferences." *PNAS* 119, no. 28 (2022).

Neilson, W. S. "Axiomatic Reference-Dependence in Behavior toward Others and toward Risk." *Economic Theory* 28, no. 3 (2006): 681-92.

Rohde, K. I. M. "A Preference Foundation for Fehr and Schmidt's Model of Inequity Aversion." *Social Choice and Welfare* 34, no. 4 (2010): 537-47.

Saito, K. "Social Preferences under Risk: Equality of Opportunity Versus Equality of Outcome." *American Economic Review* 103, no. 7 (2013): 3084-101.

Snowberg, E., and L. Yariv. "Testing the Waters: Behavior across Participant Pools." *American Economic Review* 111, no. 2 (2021): 687-719.