Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
http://www.econ.upenn.edu/pier

PIER Working Paper 08-037

"Ashamed to be Selfish"

by

David Dillenberger and Philipp Sadowski

Ashamed to be Selfish*

David Dillenberger and Philipp Sadowski†

October 13, 2008

Abstract

We study a two-stage choice problem, where alternatives are allocations between the decision maker (DM) and a passive recipient. The recipient observes choice behavior in stage two, while stage one choice is unobserved. Choosing selfishly in stage two, in the face of a fairer available alternative, may inflict shame on DM. DM has preferences over sets of alternatives that represent period two choices. We axiomatize a representation that identifies DM's selfish ranking, her norm of fairness and shame. Altruism is the most prominent motive that can explain non-selfish choice. We identify a condition under which shame to be selfish can mimic altruism, when only stage-two choice is observed by the experimenter. An additional condition implies that the norm of fairness can be characterized as the Nash solution of a bargaining game induced by the second-stage choice problem. The representation is generalized to allow for finitely many recipients and applied to a simple strategic situation, a game of trust.

JEL Classifications: C78, D63, D64, D80, D81

1. Introduction

1.1. Motivation

The notions of fairness and altruism have attracted the attention of economists in different contexts. The relevance of these motives to decision making is both intuitively convincing and well documented. For example in a classic "dictator game," where one person gets to anonymously divide, say, \$10 between herself and a partner, people tend not to take the whole amount for themselves, but to give a sum of between \$0 and \$5 to the other player. They act as if they are trading off a concern for fairness or for the other person's incremental

*We thank Roland Benabou and Wolfgang Pesendorfer for their invaluable support. We are also grateful to Eric Maskin, Stephen Morris, Charles Roddie and Tymon Tatur for helpful suggestions. This paper was written in part while the authors were graduate prize fellows at the University Center for Human Values, Princeton University. Financial support from the NSF under grant SES-0550540 is gratefully acknowledged.

†Department of Economics, University of Pennsylvania, ddill@sas.upenn.edu
Department of Economics, Duke University, p.sadowski@duke.edu

wealth and a concern for their own.[1] Thus, preferences for fairness as well as preferences for altruism have been suggested and considered (for example Fehr and Schmidt [1999], Anderoni and Miller [2002], and Charness and Rabin [2002]).

Recent experiments, however, show that this interpretation may be rash: Dana, Cain and Dawes (2006) study a variant of the same dictator game, where the dictator is given the option to exit the game before the recipient learns it is being played. If she opts out, she is given a specified amount of money and the recipient gets nothing, as the game has not taken place. It turns out that about a third of the participants choose to leave the game when offered \$9 for themselves and \$0 for the recipient. Write this allocation as (\$9, \$0). Such behavior contradicts altruistic concern regarding the recipient's payoff, because then the allocation (\$9, \$1) should be strictly preferred. It also contradicts purely selfish preferences, as (\$10, \$0) would be preferred to (\$9, \$0). Instead, people seem to suffer from behaving egoistically in a choice situation where they could dictate a fairer allocation. Hence, if they can avoid getting into such a situation, they happily do so. Real-life scenarios with this character could be:

• donating to a charity over the phone but wishing not to have been home when the call came,

• crossing the road to avoid meeting a beggar.

Our explanation of this type of behavior is the following: Whether a person's actions are observed or not plays a crucial role in determining her behavior. We term "shame" the motive that distinguishes choice behavior when observed from choice behavior when not observed. In our model, individuals are selfish when not observed. Thus, concern for another person's payoff is motivated not by altruism, but by avoiding the feeling of shame that comes from behaving selfishly when observed.[2] The interpretation is that, if people are observed, they feel shame when they do not choose the fairest available alternative.[3]

We axiomatically formalize the notion of shame and its interaction with selfishness as described above. To this end, we consider games like the one conceived by Dana et al (2006) as a two-stage choice problem. In the first stage, the decision maker (DM) chooses a "menu," a set of payoff-allocations between herself and the anonymous recipient. This choice is not observed by the recipient. In the second stage, she makes a potentially anonymous choice from the alternatives on this menu, where the recipient observes the chosen alternative in full

[1] See for example Camerer (2003).

[2] To distinguish shame from guilt, note that guilt is typically understood to involve regret, even in private, while, according to Buss (1980), *"shame is essentially public; if no one else knows, there is no basis for shame. [...] Thus, shame does not lead to self-control in private."* We adopt the interpretation that even observation of a selfish behavior without identification of its purveyor can cause shame.

[3] In a parallel work, Neilson (2006-b) entertains a very similar notion of shame. The questions and the methodology of the two works are different. Section 6 comments in more detail.

knowledge of the menu.[4] DM has well-defined preferences over sets of alternatives (menus). Our interpretation of shame as the motivating emotion allows considerations of fairness to impact preferences only through their effect on second-stage choices, where the presence of a fairer option reduces the attractiveness of an allocation. The underlying normative notion of fairness is central to our model, because assumptions on the norm of fairness are indirect assumptions on DM's preferences. Assuming a *particular* norm of fairness is difficult, descriptively as well as normatively. Instead, we impose what we consider minimal normative constraints on fairness.

Our representation results establish a correspondence between DM's norm of fairness and her choice behavior. On the one hand, this illustrates how those minimal constraints on fairness impact choice. On the other hand, the particular norm of fairness used by DM can be elicited from her choice behavior.

1.2. Illustration of Results

Denote a typical menu as $A = \{(a_1, a_2), (b_1, b_2), ...\}$, where the first and second components in each alternative are, respectively, the private payoff for DM and for the recipient. We impose axioms on DM's preferences over menus that allow us to establish a sequence of representation theorems. To illustrate our results, consider a special case of those representations:

$$U(A) = \max_{(a_1,a_2)\in A} \left[u(a_1) + \beta\varphi(a_1, a_2)\right] - \beta \max_{(b_1,b_2)\in A} \left[\varphi(b_1, b_2)\right],$$

where u and φ are increasing in all arguments. u is a utility function over private payoffs and $\varphi(a_1, a_2)$ is interpreted as the fairness of the allocation (a_1, a_2).

Alternatively, if we denote by a^* and b^* the two maximizers above, it can be written as:

$$U(A) = \underbrace{u(a_1^*)}_{\text{value of private payoff}} - \underbrace{\beta\left(\varphi(b_1^*, b_2^*) - \varphi(a_1^*, a_2^*)\right)}_{\text{shame}}.$$

This representation captures the tension between the impulse to maximize private payoff and the desire to minimize shame from not choosing the fairest alternative within a set. It evaluates a menu by the highest utility an allocation on the menu gets, where this utility depends on the menu itself. The utility function that is used to evaluate allocations is additive and has two distinct components. The first component, $u(a_1)$, gives the value of a

[4] If the exit option is chosen in the aforementioned experiment by Dana et al, as in our setup, the recipient does not observe that there was a dictator, who could have chosen another allocation. In their experiment, the recipient is further unaware that another person was involved at all. It would be interesting to see how informing the recipient that some other person had received \$9 would change the experimental findings. This would correspond to our setup.

degenerate menu (a singleton set) that contains the allocation under consideration. When evaluating degenerate menus, which leave DM with a trivial choice under observation, we assume her to be *selfish*: she prefers one allocation to another if and only if the former gives her a greater private payoff, independent of the recipient's payoff. The second component is "shame." It represents the cost DM incurs when selecting (a_1, a_2) in the face of the fairest available alternative, (b_1^*, b_2^*).

As shame is evoked whenever this fairest available alternative is not chosen, we can relate choice to a second binary relation "fairer than," which represents DM's private norm of fairness. We assume that DM's private norm of fairness induces a *Fairness Ranking* of all alternatives, which is represented by $\varphi(a_1, a_2)$. We further assume that DM's norm of fairness satisfies *Solvability*, implying that the fairness ranking is never satiated in one player's payoff, and the *Pareto* criterion in payoffs, implying that φ is increasing in all arguments.

In the special case considered here, the shame from choosing (a_1, a_2) in stage two is $\beta(\varphi(b_1^*, b_2^*) - \varphi(a_1, a_2))$. Hence, even alternatives that are not chosen may matter for the value of a set, and larger sets are not necessarily better. To see this, consider the representation above with $u(a_1) = a_1$, $\beta = \frac{1}{2}$ and $\varphi(a_1, a_2) = a_1 a_2$. Compare the sets $\{(10,1), (4,3)\}$, $\{(10,1)\}$ and $\{(4,3)\}$. Evaluating these sets we find $U\{(10,1), (4,3)\} = 9$, $U\{(10,1)\} = 10$ and $U\{(4,3)\} = 4$. To permit such a ranking, we assume a version of *Left Betweenness*, which allows smaller sets to be preferred over larger sets. Left Betweenness weakens the Set Betweenness assumption first introduced by Gul and Pesendorfer (2001), henceforth GP. Theorem 1 establishes that our weakest representation, which captures the intuition discussed thus far, is equivalent to the collection of all the above assumptions.

Selfishness leaves no room for altruism. Suppose, however, that only the second stage of the procedure is observed (for example, because DM, as in the classic dictator game, never gets to choose between menus). In this case, our representations might conform with DM behaving as if she had direct interest in the recipient's welfare and had to trade off this altruistic motive with concerns about her private payoff. We argue that it is hard to reconcile such an interpretation with observing any choice reversal in stage two. Thus, when observing stage two in isolation, shame can mimic altruism only if the induced choice ranking is set independent. Theorem 2 establishes that, given the assumptions made so far, an additional separability assumption on preferences over sets, *Consistency*, is equivalent to the existence of such a ranking. In the special case of our representation considered above, the induced choice behavior satisfies *Consistency*. To see this, regroup the terms as follows:

$$U(A) = \underbrace{\max_{(a_1,a_2)\in A} [u(a_1) + \beta\varphi(a_1, a_2)]}_{\text{second stage choice criterion}} - \underbrace{\beta \max_{(b_1,b_2)\in A} [\varphi(b_1, b_2)]}_{\text{effect of fairest alterative}}.$$

We further specify the norm of fairness by assuming that the private payoffs to the two players have *Independent Fairness Contributions*: The fairness contribution of raising one player's payoff can not depend on the level of the other player's payoff. The idea is that interpersonal utility comparisons are infeasible. With this additional assumption, Theorem 3 establishes that there are two utility functions, v_1 and v_2, evaluated in the payoff to DM and the recipient respectively, such that the value of their product represents the fairness ranking, $\varphi(a_1, a_2) = v_1(a_1) v_2(a_2)$. Thus, the fairest alternative within a set of alternatives can be characterized as the Nash Bargaining Solution (NBS) of an associated game. Because the utility functions used to generate this game are private, so is the norm.[5] We argue that when based on true selfish utilities, the NBS is a convincing fairness criterion in our context. Those utilities, however, may not be publicly known, especially in anonymous choice situations, and therefore, DM may not be able to base her evaluation on true selfish utilities. Nevertheless, one can assess the descriptive appeal of the representation by asking whether the utilities comprising the norm at least resemble selfish utilities.

Example: Let $u(a_1) = a_1$, $\varphi(a_1, a_2) = v_1(a_1) v_2(a_2) = a_1 a_2$ and $\beta = \frac{1}{2}$. This implies that selfish utility u is risk neutral and unbounded, and that the utilities v, which are used to generate the fairness ranking, coincide with u. Shame is half the difference between the Nash-product of the fairest and the chosen alternatives. Reconsider the experiment by Dana et al (2006) mentioned above, with the added constraint that only integer values are possible allocations. The set $A = \{(10, 0)(9, 1)(8, 2), ..., (0, 10)\}$ then corresponds to the dictator game. It induces the imaginary bargaining game with possible utility-allocations $\{(10, 0)(9, 1)(8, 2), ..., (0, 10), (0, 0)\}$, where the imaginary disagreement point is $\lim_{(x,y)\to 0} \left(v_1^{-1}(x), v_2^{-1}(y)\right) = (0, 0)$. According to the NBS, $(5, 5)$ would be the outcome of the bargaining game. Its fairness is $5 \cdot 5 = 25$. To trade off shame with selfishness, DM chooses the alternative that maximizes the sum of private utility and fairness, $a_1 + a_1 a_2$, which is $(6, 4)$. Its fairness is $6 \cdot 4 = 24$ and the shame incurred by choosing it is $\frac{1}{2}$. Hence $U(A) = 5.5$. From the singleton set $B = \{(9, 0)\}$, which corresponds to the exit option in the experiment, the choice is trivial and $U(B) = 9$. This example illustrates both the tension DM is exposed to when choosing from a large set and the reason why she might prefer a smaller menu.

The organization of the paper is as follows: Section 2 presents the basic model and a representation that captures the concepts of fairness and shame. Section 3 isolates a choice criterion from the choice situation. Section 4 further specifies the fairness ranking. Section

[5] Therefore, the fairness ranking could also be represented by a different functional, based on different utilities.

5 extends the representation to finitely many other players and suggests an application to a simple strategic situation, a game of trust. Section 6 concludes by pointing out connections to existing literature.

2. The Model

Let K be the set of all finite subsets of $\mathbb{R}^2_+$.[6] Any element $A \in K$ is a finite set of alternatives. A typical alternative $\mathbf{a} = (a_1, a_2)$ is interpreted as a payoff pair, where a_1 is the private payoff for DM and a_2 is the private payoff allocated to the (potentially anonymous) other player, the recipient. Endow K with the topology generated by the Hausdorff metric, which is defined for any pair of non-empty sets, $A, B \in K$, by:

$$d_h(A, B) := \max \left[\max_{\mathbf{a} \in A} \min_{\mathbf{b} \in B} d(\mathbf{a}, \mathbf{b}), \max_{\mathbf{b} \in B} \min_{\mathbf{a} \in A} d(\mathbf{a}, \mathbf{b}) \right],$$

where $d : \mathbb{R}^2_+ \to \mathbb{R}_+$ is the standard Euclidian distance.

Let $\succ$ be a continuous preference relation (weak order) over K. We write $A \succ B$ if DM strictly prefers A to B. The associate weak preference, $\succeq$ and the indifference relation, $\sim$ are defined in the usual way.

The choice of a menu $A \in K$ is not observed by the recipient, while the choice *from* any menu is. We call the impact this observation has on choice "shame." Of course various other regarding preferences that are not impacted by observation could be present as well. We do not account for those, as our aim is not to describe a range of possible attitudes toward others, but to derive a tractable representation according to which DM distinguishes the two stages in an intuitive way.

The first axiom specifies DM's preferences over singleton sets.

P_1 **(Selfishness)** $\{\mathbf{a}\} \succ \{\mathbf{b}\}$ *if and only if* $a_1 > b_1$.

A singleton set $\{\mathbf{a}\}$ is a degenerate menu that contains only one feasible allocation, (a_1, a_2). It leaves DM with a trivial choice to be made when being observed in the second stage. Therefore, the ranking over singleton sets can be thought of as the ranking over allocations that are imposed on DM. We contend that there is no room for shame in this situation; choosing between two singleton sets reveals DM's "true" preferences over allocation outcomes. The axiom states that DM is not concerned about the payoff to the second player when evaluating such sets; she compares any pair of alternatives based solely on the first

[6] With $\mathbb{R}_+$ we denote the positive reals including 0. $\mathbb{R}_{++}$ denotes the positive reals without 0.

component, her private payoff. If, for example, DM had an altruistic concern for fairness in the dictator game previously described, she would strictly prefer the menu $\{(9,1)\}$ to $\{(9,0)\}$. P_1 rules out such altruistic concerns. Negative emotions regarding the other player, such as spite or envy, are ruled out as well.

The next axiom captures the idea that shame is a mental cost, which is invoked by unchosen alternatives.

P_2 (**Strong Left Betweenness**) *If $A \succeq B$, then $A \succeq A \cup B$. Further, if $A \succ B$ and $\exists C$ such that $A \cup C \succ A \cup B \cup C$, then $A \succ A \cup B$.*

We assume that adding unchosen alternatives to a set can only increase shame. Therefore, no alternative is more appealing when chosen from $A \cup B$, than when chosen from one of the smaller sets, A or B. Hence, $A \succeq B$ implies $A \succeq A \cup B$.[7] Furthermore, if additional alternatives add to the shame incurred by the original choice from a menu $A \cup C$, then they must also add to the shame incurred by any choice from the smaller menu A. Thus, if there is C such that $A \cup C \succ A \cup B \cup C$ and if $A \succ B$, then $A \succ A \cup B$.

Shame, which is the only motive DM knows beyond selfishness, must refer to some personal norm that determines what the appropriate choice should have been. In our interpretation, this norm is to choose one of the fairest available allocations. Interpreting "fairness" as a property of an allocation, which is independent of the menu it is on, we consider a binary relation $\succ_f$ over $\mathbb{R}^2_+$ as a second primitive.

Definition: If $\mathbf{b} \succ_f \mathbf{a}$, we say that DM considers $\mathbf{b}$ to be *fairer than* $\mathbf{a}$.

Some of the axioms below are imposed on $\succ_f$ rather than on $\succ$ and are labeled by F instead of P. The underlying notion of fairness is at the heart of those assumptions.[8] To make them descriptively intuitive, we emphasize their normative appeal, implying that DM will want her norm of fairness to satisfy them. Making these assumptions directly on $\succ_f$ is natural. The relation $\succ_f$ is not directly observable, but the next axiom relates it to observable choice behavior. One contribution of our work is that the implications of F-axioms on $\succ$ are most easily understood from the representation.

[7] This is the "Left Betweenness" axiom. It appears in Dekel, Lipman and Rustichini (2005) and is a weakening of "Set Betweenness" as first posed in GP.

[8] In everyday language, "fair" is sometimes used to capture various different notions. According to the Merriam-Webster Collegiate Dictionary (Tenth Edition, 2001) "*Fair implies an elimination of one's own feelings, prejudices, and desires so as to achieve a proper balance of conflicting interests.*" This is the definition of "fair" we base our arguments on.

P_3 (**Shame**) *If* $\exists A \in K$ *with* $\mathbf{a} \in A$, *such that* $A \succ A \cup \{\mathbf{b}\}$, *then* $\mathbf{b} \succ_f \mathbf{a}$.[9]

$A \succ A \cup \{\mathbf{b}\}$ implies that $\mathbf{b}$ adds to the shame incurred by the original choice in A. The interpretation is that DM is concerned about not choosing one of the fairest available alternatives. Thus, $\mathbf{b}$ must be fairer than any alternative in A, in particular $\mathbf{b} \succ_f \mathbf{a}$.

Definition: We say that DM is *susceptible to shame* if there exists A and B with $A \succ A \cup B$.

F_1 (**Fairness Ranking**) $\succ_f$ *is an anti-symmetric and negatively transitive binary relation.*

Our discussion rests on the assumption that DM can rank alternatives according to their fairness. In $\mathbb{R}^2_+$ and with increasing utility from self-payoffs, this assumption is not unreasonably restrictive.[10]

Combined with P_3, F_1 implies that only one alternative in each menu, the fairest, is responsible for shame.

F_2 (**Pareto**) *If* $\mathbf{a} \geq \mathbf{b} > \mathbf{0}$ *and* $\mathbf{a} \neq \mathbf{b}$, *then* $\mathbf{a} \succ_f \mathbf{b}$.

According to this axiom, absolute, as opposed to relative, well-being matters; the Pareto criterion excludes notions such as "strict inequality aversion." The resulting concept of fairness must have some concern for efficiency. In the case where there truly is no potential for redistribution, we believe that people find the Pareto criterion a reasonable requirement for one allocation to be fairer than another.[11]

[9] The notion of "fairer than" is analogous to the definition of "more tempting than" in Gul and Pesendorfer (2005).

[10] If, instead, there were a globally most prefered self-payoff, this assumption would rule out very reasonable preference rankings.

[11] In many contexts, people would disagree with the statement that the allocation $(1million, 6)$ is fairer than $(5, 5)$. On the basis of the definition in footnote 10, however, we claim that the opposition to $(1million, 6)$ as a fair allocation can only be based on the implicit premise that there must be some mechanism to divide the gains more evenly (Such a mechanism would imply the availability of a third option, which would render both of the above allocations unfair.) In an explicit choice situation this premise cannot be sustained. The Pareto property has indeed been advocated in the philosophical literature on fairness. Rawls (1971), for example, proposes the idea of "original position," a mental exercise whereby a group of rational people must establish a principle of fairness (e.g. when distributing income) without knowing beforehand where on the resulting pecking order they will end up themselves. Requiring that the allocation satisy *Pareto* makes much sense in such an environment.

F_3 **(Solvability)** *If* $(a_1, 0) \nprec_f (b_1, b_2)$ *then* $\exists x$ *such that* $(a_1, x) \sim_f (b_1, b_2)$*. Analogously, if* $(0, a_2) \nprec_f (b_1, b_2)$ *then* $\exists y$ *such that* $(y, a_2) \sim_f (b_1, b_2)$*.*

Ignoring the qualifier, the axiom states that in order to make two allocations deemed equally fair, any variation in the level of one person's payoff can always be compensated by appropriate variation in the level of the other person's payoff. This requires $\succ_f$ never to be satiated in any person's payoff. Relying on F_2, the qualifiers take into account that monetary payoffs are bounded below by 0. For example, F_3 implies that there is a sum x, such that $(x, 1) \sim_f (10, 10)$. This assumption captures the insight that any fairness ranking with a concern for efficiency must go beyond the Pareto principle and trade off, in some manner, payoffs across individuals.

As $\succ$ is continuous, $\succ_f$ is continuous in all alternatives for which P_3 relates $\succ$ to $\succ_f$. $F_1 - F_3$ imply that this is the case on $\mathbb{R}_+ \times \mathbb{R}_{++}$.[12] Assuming that $\succ_f$ is continuous even in alternatives for which P_3 does not relate $\succ$ to $\succ_f$ has obviously no implication for choice. For ease of exposition, we assume in all what follows that $\succ_f$ is continuous on all of $\mathbb{R}^2_+$.

Theorem 1 *If DM is susceptible to shame, then* $\succ$ *and* $\succ_f$ *satisfy* $P_1 - P_3$ *and* $F_1 - F_3$ *respectively, if and only if there exist continuous and strictly increasing functions* $u : \mathbb{R}_+ \to \mathbb{R}$ *,* $\varphi : \mathbb{R}^2_+ \to \mathbb{R}$ *and a continuous function* $g : \mathbb{R}^2_+ \times \varphi\left(\mathbb{R}^2_+\right) \to \mathbb{R}$*, weakly increasing in its second argument and satisfying:* $g(\mathbf{a}, x) \gtreqless 0$ *whenever* $\varphi(\mathbf{a}) \lesseqgtr x$*, such that the function* $U : K \to \mathbb{R}$ *defined as* $U(A) = \max\limits_{\mathbf{a}\in A}\left[u(a_1) - g\left(\mathbf{a}, \max\limits_{\mathbf{b}\in A}\varphi(\mathbf{b})\right)\right]$ *represents* $\succ$ *and* φ *represents* $\succ_f$*.*
If DM is not susceptible to shame, $g \equiv 0$*.*

All detailed proofs are in the appendix. We now highlight the important steps. As both $\succ$ and $\succ_f$ are continuous binary relations, they can be represented by continuous functions $U : K \to \mathbb{R}$ and $\varphi : \mathbb{R}^2_+ \to \mathbb{R}$ respectively. φ is an increasing function as implied by *Pareto* (F_2). The combination of *Strong Left Betweenness* (P_2), *Shame* (P_3) and *Fairness Ranking* (F_1) implies GP's *Set Betweenness* (SB) property: $A \succeq B$ implies $A \succeq A \cup B \succeq B$. GP demonstrate that imposing SB on preferences over sets makes every set indifferent to a certain subset of it, which includes at most two elements (Lemma 2 in their paper). Hence we confine our attention to a subset of our domain that includes all sets with cardinality no greater than 2. *Selfishness* (P_1) and P_3 imply that a set $\{\mathbf{a}, \mathbf{b}\}$ is strictly inferior to $\{\mathbf{a}\}$ if and only if $a_1 > b_1$ and $\mathbf{b} \succ_f \mathbf{a}$. We can then strengthen GP's Lemma 2 and state that any set is indifferent to some two-element set that includes one of the fairest allocations in the

[12] $\succ_f$ is relevant for choice in alternative $\mathbf{b}$, if and only if there is $\mathbf{c}$ with $\mathbf{c} \prec_f \mathbf{b}$ and $c_1 > b_1$, which requires $c_2 < b_2$. Thus $b_2 > 0$ is necessary for the construction of $\mathbf{c}$.

original (larger) set. Using *Solvability* (F_3) we show the continuity of the second component, the function g, in the representation.

The representation in Theorem 1 highlights the basic trade-off between private payoff and shame as the only concepts DM may care about. There are at most two essential alternatives within a set, to be interpreted as the "chosen" and the "fairest" alternative, **a** and **b** respectively. For the latter, its fairness, $\varphi(\mathbf{b})$, is a sufficient statistic for its impact on the set's value. DM suffers from shame, measured by $g(\mathbf{a}, \varphi(\mathbf{b}))$, whenever $\varphi(\mathbf{a}) < \varphi(\mathbf{b})$, where $\varphi(\mathbf{a})$ is the fairness of the chosen alternative. The representation captures the idea of shame being an emotional cost that emerges whenever the fairest available allocation is not chosen. Its magnitude may depend on the fairness of the chosen allocation.

The main contribution of Theorem 1 is that it determines when DM's fairness ranking, $\succ_f$, can be elicited from choice behavior: all functions in the representation are continuous and hence, for $\mathbf{a}, \mathbf{b} \in \mathbb{R}_+ \times \mathbb{R}_{++}$ and $\mathbf{b} \succ_f \mathbf{a}$, there is $\mathbf{c}$ such that $U(\{\mathbf{a}, \mathbf{c}\}) > U(\{\mathbf{a}, \mathbf{b}, \mathbf{c}\})$, implying $\varphi(\mathbf{b}) > \varphi(\mathbf{a})$. As the function φ is continuous, it is uniquely determined by choice behavior on its entire domain, $\mathbb{R}^2_+$, if the axioms $P_1 - P_3$ and $F_1 - F_3$ hold.

Note that the properties of the function g and the max operator inside imply that the second term is always a cost (non-positive). The other max operator implies that DM's payoff will never lie below b_1, which is her payoff as suggested by the fairest allocation. Thus, any deviations by DM from choosing the fairest allocation will be in her own favor. These observations justify labeling said cost as "shame."

From the representation, it is easy to see that the induced choice correspondence,

$$C(A) := \left\{ \underset{\mathbf{a} \in A}{\arg\max} \left[u(a_1) - g\left(\mathbf{a}, \max_{\mathbf{b} \in A} \varphi(\mathbf{b}) \right) \right] \right\}$$

may be context dependent in the sense that a higher degree of shame may affect choice. In other words, if we define a binary relation "better choice than," $\succ_c$, by $\mathbf{a} \succ_c \mathbf{b}$ if $\exists B$ with $\mathbf{b} \in B$, such that $B \cup \{\mathbf{a}\} \succ B$, then this binary relation need not be acyclic. This feature may be plausible when shame is taken into account. In the next section we spell out the implications of enforcing a context-independent criterion for choice.

3. A Second-Stage Choice Ranking

In many situations, only second-stage choice may be observable. For example, the standard dictator game corresponds only to second-stage choice in our setup. Typical behavior in various versions of this game, where subjects tend to give part of the endowment to the recipient, is often interpreted as motivated by an altruistic motive. We interpret altruism to imply that the recipient's welfare is a good, just as selfishness implies that DM's private payoff is a good.[13] If DM had those two motives, she would have to make a trade-off between them. As in the case of two generic goods, very basic assumptions would lead to a context-independent choice ranking of alternatives. As we point out at the end of section 2, we can define a binary relation "better choice than," $\succ_c$, by $\mathbf{a} \succ_c \mathbf{b}$ if $\exists B$ with $\mathbf{b} \in B$, such that $B \cup \{\mathbf{a}\} \succ B$. This binary relation need not be acyclic: Different choice problems, A and B, may lead to different second-stage rankings of $\mathbf{a}$ and $\mathbf{b}$, for $\mathbf{a}, \mathbf{b} \in A \cap B$. If no cycles occur, second-stage behavior might look as if it were generated by, for instance, a trade-off between selfishness and altruism, even though observation of stage-one choice would rule this out. If, on the other hand, cycles are observed in stage-two choice, simple altruistic motives cannot be solely responsible for behavior that is not purely selfish. In this section we identify a condition on preferences that makes DM's second-stage choice independent of the choice set. This implies finding a function $\psi : \mathbb{R}^2_+ \to \mathbb{R}$ that assigns a value to each $\mathbf{a} \in A$, such that $\mathbf{a}$ is a choice from A only if $\psi(\mathbf{a}) \geq \psi(\mathbf{b})$ for all $\mathbf{b} \in A$.

Definition: $X := \{(\mathbf{a}, \mathbf{b}) : \{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}\}$ is the set of all pairs of alternatives generating strict *Set Betweenness*.

For any set of two allocations $\{\mathbf{a}, \mathbf{b}\}$, we interpret the preference ordering $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$ as an indication of a discrepancy between what DM chooses ($\mathbf{a}$) and the alternative she deems to be the fairest ($\mathbf{b}$), which causes her choice to bear shame. This shame, however, is not enough to make her choose $\mathbf{b}$.

Combined with F_1, Shame (P_3) implies that choice between sets depends on the fairness of the fairest alternative in the set. The next axiom relates choice to the fairness of the chosen alternative as well: The fairer DM's choice, the less shame she feels.

P_4 **(Fairer is Better)** *If for* $\{\mathbf{a}\} \sim \{\mathbf{a}'\}$ *we have* $\{(\mathbf{a}, \mathbf{b}), (\mathbf{a}', \mathbf{b})\} \subseteq X$ *and* $\mathbf{a} \succ_f \mathbf{a}'$, *then* $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}', \mathbf{b}\}$.

Axiom P_4 implies that only the fairness of the chosen alternative matters for its impact on shame.

Given $P_1 - P_4$ and $F_1 - F_3$, an additional separability assumption is equivalent to separable shame, and thus to a set-independent choice ranking.

[13] This interpretation is based on the following definition of altruism (Merriam-Webster Collegiate Dictionary [Tenth Edition, 2001]): "*Unselfish regard for or devotion to the welfare of others.*" We understand this definition as ruling out any considerations that condition on available but unchosen alternatives.

P_5 (**Consistency**) *If*

$$\{(\mathbf{a},\mathbf{b}),(\mathbf{a},\mathbf{d}),(\mathbf{a}',\mathbf{b}'),(\mathbf{a}',\mathbf{d}'),(\mathbf{c},\mathbf{b}),(\mathbf{c}',\mathbf{b}'),(\mathbf{c},\mathbf{d}),(\mathbf{c}',\mathbf{d}')\}\subseteq X,$$

then $\{\mathbf{a},\mathbf{b}\}\sim\{\mathbf{a}',\mathbf{b}'\}$ *and* $\{\mathbf{a},\mathbf{d}\}\sim\{\mathbf{a}',\mathbf{d}'\}$ *imply* $\{\mathbf{c},\mathbf{b}\}\succ\{\mathbf{c}',\mathbf{b}'\}\Leftrightarrow\{\mathbf{c},\mathbf{d}\}\succ\{\mathbf{c}',\mathbf{d}'\}$.

We make no claim about the normative or descriptive appeal of this assumption. Instead, we view it as an empirical criterion: If the condition is not met, observation of stage-two choice should suffice to distinguish altruism from shame as the motive behind DM's other-regarding behavior. The axiom requires independence between the impact of the chosen and the fairest alternative on the set ranking;

$$\{(\mathbf{a},\mathbf{b}),(\mathbf{a},\mathbf{d}),(\mathbf{a}',\mathbf{b}'),(\mathbf{a}',\mathbf{d}'),(\mathbf{c},\mathbf{b}),(\mathbf{c}',\mathbf{b}'),(\mathbf{c},\mathbf{d}),(\mathbf{c}',\mathbf{d}')\}\subseteq X$$

implies that from each of the sets $\{\mathbf{a},\mathbf{b}\},\{\mathbf{a},\mathbf{d}\},\{\mathbf{a}',\mathbf{b}'\},\{\mathbf{a}',\mathbf{d}'\},\{\mathbf{c},\mathbf{b}\},\{\mathbf{c}',\mathbf{b}'\},\{\mathbf{c},\mathbf{d}\}$ and $\{\mathbf{c}',\mathbf{d}'\}$, the alternative listed first is chosen in the second stage despite the availability of a fairer alternative, which is listed second. Assume, without loss of generality that $\{\mathbf{a}\}\succ\{\mathbf{a}'\}$. Suppose there are two pairs of fairer and less attractive alternatives, $\mathbf{b},\mathbf{b}'$ and $\mathbf{d},\mathbf{d}'$, such that for each of them pairing their members with $\mathbf{a}$ and $\mathbf{a}'$, respectively, gives rise to indifference. In the context of Theorem 1, this implies that both pairs induce the same shame differential, which exactly cancels the selfish preference of $\{\mathbf{a}\}$ over $\{\mathbf{a}'\}$: $\{\mathbf{a},\mathbf{b}\}\sim\{\mathbf{a}',\mathbf{b}'\}$ and $\{\mathbf{a},\mathbf{d}\}\sim\{\mathbf{a}',\mathbf{d}'\}$. Then, the axiom states that pairing the members of $\mathbf{b},\mathbf{b}'$ or $\mathbf{d},\mathbf{d}'$ with any other chosen alternatives $\mathbf{c}$ and $\mathbf{c}'$, respectively, must also lead to the same differential in shame. In particular, $\{\mathbf{c},\mathbf{b}\}\succ\{\mathbf{c}',\mathbf{b}'\}$ implies $\{\mathbf{c},\mathbf{d}\}\succ\{\mathbf{c}',\mathbf{d}'\}$. Again, the validity of this technical assumption in a given context is an empirical question.

Theorem 2 *If DM is susceptible to shame, then $\succ$ and $\succ_f$ satisfy P_1-P_5 and F_1-F_3 respectively, if and only if there exist continuous and strictly increasing functions $u:\mathbb{R}_+\to\mathbb{R}$ and $\varphi:\mathbb{R}_+^2\to\mathbb{R}$, such that the function $U:K\to\mathbb{R}$ defined as*

$$U(A)=\max_{\mathbf{a}\in A}\left[u(a_1)+\varphi(a_1,a_2)\right]-\max_{b\in A}\left[\varphi(b_1,b_2)\right]$$

represents $\succ$ and φ represents $\succ_f$.

The proof constructs a path in the (a_1,a_2)-plane such that the fairness $\varphi(\mathbf{a})$ increases along this path. Then, on two neighboring indifference curves in the $(\mathbf{a},\varphi(\mathbf{b}))$-space, $\varphi(\mathbf{b})$

increases, as $\mathbf{a}$ varies along the path. Relying on P_5, these indifference curves allow us to rescale $\varphi(\mathbf{b})$ to make the representation of $\succ$ quasi-linear.[14] Separability is then immediate. Since the proof of Theorem 2 is a special case of the proof of Theorem 4, we only go through the more general case in detail in the appendix.

The representation isolates a choice criterion that is independent of the choice problem: DM's behavior is governed by maximizing

$$u(a_1) + \varphi(a_1, a_2).$$

The value of the set is reduced by

$$\max_{b \in A} \varphi(b_1, b_2),$$

a term that depends solely on the fairest alternative in the set. Grouping the terms differently reveals the trade-off between self-payoff, $u(a_1)$, and the shame involved with choosing $\mathbf{a}$ from the set A:

$$\max_{\mathbf{b} \in A} [\varphi(b_1, b_2) - \varphi(a_1, a_2)] \geq 0.$$

Note that now shame takes an additively separable form, depends only on the fairness of both alternatives, and is increasing in the fairness of the fairest and decreasing in that of the chosen alternative. If $P_1 - P_4$ and $F_1 - F_3$ hold, then, according to Theorem 2, P_5 is equivalent to having a set-independent choice ranking.

4. Specifying a Fairness Ranking

In this section we impose one more axiom on $\succ_f$ to further characterize the fairness ranking. It asserts that the fairness contribution of one person's marginal payoff cannot depend on the initial payoff levels.

F_4 (**Independent Fairness Contributions**) *If* $(a_1, a_2) \sim_f (b_1, b_2)$ *and* $(a_1', a_2) \sim_f (a_1, b_2) \sim_f (b_1, b_2')$, *then* $(a_1', b_2) \sim_f (a_1, b_2')$.

The axiom is illustrated in figure 1. If $a_1 = a_1'$ or $b_2 = b_2'$, this axiom is implied by F_1, F_2 and the continuity of $\succ_f$. For $a_1 \neq a_1'$ and $b_2 \neq b_2'$, the statement is more subtle. Consider first a stronger assumption:

F_4' (**Strong Independent Fairness Contributions**) $(a_1, a_2) \sim_f (b_1, b_2)$ *and* $(a_1', a_2) \sim_f (b_1, b_2')$ *imply* $(a_1', b_2) \sim_f (a_1, b_2')$.

[14] A more elaborate discussion on this technique appears after Theorem 3.

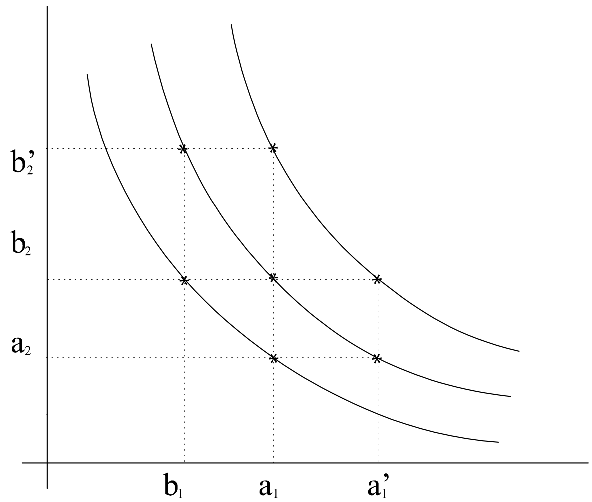

Figure 1: Independent Fairness Contributions.

The fairness contribution of one person's marginal payoff cannot depend on the initial payoff level of the other person: It is unclear to DM how much an increase in monetary payoff means to the recipient, because even if the (marginal) utility of the recipient were known to DM, she could not compare it to her own, as interpersonal utility comparisons are infeasible. The qualifier in F_4' establishes that DM considers the fairness contribution of changing her own payoff from a_1 to a_1' given the allocation (a_1, a_2) to be the same as that of changing the recipient's payoff from b_2 to b_2' given (b_1, b_2). F_4' then states that starting from the allocation (a_1, b_2), changing a_1 to a_1' should again be as favorable in terms of fairness as changing b_2 to b_2'. This is the essence of *Independent Fairness Contributions.* The stronger qualifier $(b_1, b_2') \sim_f (a_1, b_2) \sim_f (a_1', a_2)$ in F_4 weakens the axiom. For example, the fairness ranking $(a_1, a_2) \succ_f (b_1, b_2)$ if and only if $\min(a_1, a_2) > \min(b_1, b_2)$ is permissible under F_4, but not under F_4'.[15]

Theorem 3 $\succ_f$ *satisfies* $F_1 - F_4$, *if and only if there are continuous, increasing and unbounded functions* $v_1, v_2 : \mathbb{R}_+ \to \mathbb{R}_{++}$, *such that* $\varphi(\mathbf{a}) = v_1(a_1)\, v_2(a_2)$ *represents* $\succ_f$.

Luce and Tukey (1964) prove the necessity and sufficiency of *Solvability* and the *Corresponding Trade-offs Condition* (the label they use for F_4) to admit an additive representation. To show how a proof works, we repeatedly use axiom F_4 to establish that if

[15] F_4 is refered to as the *Hexagon condition* or the *Corresponding Trade-offs Condition* (Keeney and Raiffa [1976]), F_4' as the *Thomsen condition.* With F_2 and F_3, F_4' is implied by F_4. See Karni and Safra (1998) for a proof.

$(a_1, a_2) \sim_f (a_1', a_2')$ and $(a_1, \widetilde{a}_2) \sim_f (a_1', \widetilde{a}_2')$, then $(\widetilde{a}_1, a_2) \sim_f (\widetilde{a}_1', a_2') \Leftrightarrow (\widetilde{a}_1, \widetilde{a}_2) \sim_f (\widetilde{a}_1', \widetilde{a}_2')$. With this knowledge, we can create a monotone increasing mapping $a_2 \to \gamma(a_2)$ that transforms the original indifference map to be quasi-linear with respect to the first coordinate in the $(a_1, \gamma(a_2))$ plane. Keeney and Raiffa (1976) refer to the procedure we employ as the lock-step procedure. Quasi-linearity implies that there is an increasing continuous function $\xi : \mathbb{R}_+ \to \mathbb{R}$, such that $\varphi(\mathbf{a}) := \xi(a_1) + \gamma(a_2)$ represents $\succ_f$. Define $v_1(a_1) := \exp(\xi(a_1))$ and $v_2(a_2) := \exp(\gamma(a_2))$. Then $v_1, v_2 : \mathbb{R}_+ \to \mathbb{R}_{++}$ are increasing and continuous and if we redefine $\varphi(\mathbf{a}) := v_1(a_1) v_2(a_2)$, it represents $\succ_f$.

This representation suggests an appealing interpretation of the fairness ranking DM is concerned about: She behaves *as if* she had in mind two increasing and unbounded utility functions, one for herself[16] and one for the recipient. By mapping the alternatives within each set into the associated utility space, any choice set induces a finite bargaining game where only the disagreement point is unspecified. DM then identifies the fairest alternative within a set *as if* she also had in mind a disagreement point, that makes this alternative the Nash Bargaining Solution[17] of the game.[18] Moreover, the fairness of all alternatives can be ranked according to the same functional, namely the Nash product.

Remember that F_3 requires trading off marginal payoffs. The tension of having to trade off marginal payoffs without being able to compare their welfare contribution (F_4) is common in a range of social-choice problems.[19] Our axioms are weak in the sense that they do not constrain DM in this trade-off, as long as she takes into account that the fairness contribution of increasing one person's payoff should not depend on the other's payoff. The power of Theorem 3 is that it bases a representation on these weak assumptions. The downside is that the form of this representation is not unique, as the utilities v_1 and v_2 are not observable independent of the norm of fairness. For example, there is another pair of increasing utility functions such that DM is concerned about their sum, that is, she acknowledges efficiency as the only fairness criterion.

To underline the appeal of the Nash product as a descriptive representation of fairness,[20]

[16] This utility function need not agree with her true utility for personal payoffs, u. The interpretation is that DM is concerned about the recipient's perception of her choice. The recipient, however, may not know DM's true utility, especially under anonymity.

[17] See Nash (1950).

[18] The imaginary disagreement point is determined by $\lim_{(x,y)\to 0} \left(v_1^{-1}(x), v_2^{-1}(y)\right)$. It could be some finite and weakly positive pair of utility payoffs. In particular it could be $(0,0)$, which corresponds to DM imagining that players walk away in the case that no agreement is reached. It could also be negative. This corresponds to DM imagining that players have an extra incentive to find an agreement: there is a cost to disagreement.

[19] For a review, see Hammond (1990).

[20] Even though u and v_1 do not have to agree, our interpretation might be more convincing when they resemble each other empirically. In particular it is more appealing if DM's actual utility from self-payoff u is unbounded.

we now point out how DM might reason within the constraints of the axioms:

We justified the Pareto criterion, F_2, as a plausible axiom for the fairness ranking. As argued above, concern for fairness requires the acknowledgment of some form of interpersonal comparability of preferences' intensity. If utilities were known cardinally, symmetry in terms of utility payoffs is the other criterion we would expect the ranking to satisfy.[21] In our context, this implies independence of the role people play, dictator or recipient. However, utilities are inherently ordinal, rendering such a comparison infeasible. At best we can, if we assume people to have cardinal utilities that reflect their attitudes toward risk, determine marginal utilities up to scaling. Mariotti (1997), for example, considers a context in which *"interpersonal comparisons of utility are meaningful; that is, there exists an (unknown) rescaling of each person's utility which makes utilities interpersonally comparable."* At the same time, however, *"interpersonal comparisons of utility are not feasible."* Assume there is a correct interpersonal utility scaling, but DM cannot determine it. Can she guarantee that for this unknown scaling both symmetry and Pareto are satisfied? They would have to be satisfied for all potential scalings. Mariotti establishes that the NBS is the only criterion with this property.

Even more appealing is an interpretation of the NBS as the fairest allocation that is related to Gauthier's (1986) principle of "moral by agreement": Trying to assess what is fair, but finding herself unable to compare utilities across individuals, DM might refer to the prediction of a symmetric mechanism for generating allocations. In particular, DM might ask what would be the allocation if both she and the recipient were to bargain over the division of the surplus. To answer this question, she does not need to assume the intensities of the two preferences. This is a procedural interpretation that is not built on the axioms: DM is not ashamed of payoffs, but of using her stronger position in distributing the gains. It is, then, the intuitive and possibly descriptive appeal of the NBS in many bargaining situations that makes it normatively appealing to DM in our context.[22] Theorem 3 establishes the behavioral equivalence of this interpretation and our axioms.

The Pareto and the Solvability axioms, F_2 and F_3 respectively, rule out fairness rankings with $(x, 0) \sim_f (0, y)$ for all x, y. In particular the Nash product with linear utility functions v_1, v_2 is ruled out as a criterion for fairness. Such orderings could easily be accommodated by posing *Pareto* and *Solvability* only on $\mathbb{R}^2_{++}$. As a consequence, φ would be strictly increasing only on $\mathbb{R}^2_{++}$ and v_1, v_2 would only have to be weakly positive, $v_1, v_2 : \mathbb{R}_+ \to \mathbb{R}_+$.[23] These

[21] This reasoning leads Rawls (1971) to suggest *Pareto* and *Symmetry* as the two criteria a decision maker under a veil of ignorance should respect.

[22] The descriptive value of the NBS has been tested empirically. For a discussion see Davis and Holt (1993) pages 247-55. Further, multiple seemingly natural implementations of it have been proposed (Nash [1953] , Osborne and Rubinstein [1994]).

[23] As can be seen in the proof of Theorem 2, this would imply the possibility of $(-\infty, -\infty)$ as an imaginary

weaker axioms would still rule out the Maximin as a criterion for fairness.

Remark: Any concern DM has about fairness originates from being observed. Consequently, DM should expect a potentially anonymous observer to share her notion of what is fair: Her private norm of fairness, which we observe indirectly, should reflect her concern about not violating a social norm. If the observed choice situation is anonymous, DM does not know the recipient's identity and is aware that the recipient does not know hers. Therefore, the ranking cannot depend on either identity. Combining this with the idea that fairness of an allocation should not depend on the role a person plays, whether dictator or recipient, one might want to impose symmetry of the fairness ranking in terms of direct payoffs.

F_5 **(Symmetry)** $(a_1, a_2) \sim_f (a_2, a_1)$.

Adding this assumption constrains $v_1(a) = v_2(a)$ in the representation of Theorem 3. The numerical example given in the introduction features the combination of Theorem 2 and Theorem 3, where all functions involved are the identity. For brevity, we will not repeat it here.

5. Extensions

5.1. Multiple Recipients

The underlying idea is that DM (without loss of generality individual 1) is concerned about $N-1 \geqslant 2$ other individuals, whose payoffs depend on her choice. In analogy to section 2, let K be the set of all finite subsets of $\mathbb{R}_+^N$. Any element $A \in K$ is a finite set of alternatives. A typical alternative $\mathbf{a} = (a_1, a_2, ..., a_N)$ is interpreted as a payoff vector, where a_n is the payoff allocated to individual n. We write, for example, $(a_m, a_n, \mathbf{a}_{-m,n})$ as the alternative with payoff a_m to individual m, payoff a_n to individual n and $\mathbf{a}_{-m,n} \in \mathbb{R}_+^{N-2}$ lists all other individuals' payoffs in order. We endow K with the topology generated by the Hausdorff metric.

Let $\succ$ be a continuous preference relation over K. Most of the axioms we impose on $\succ$ in section 2 can be readily applied to $\succ$ on this new domain. We define $\succ_f$ in analogy to the previous definition. Instead of F_3 we write

F_3^N **(Weak Solvability)** *If* $(a_n, \mathbf{0}) \not\succ_f \mathbf{b}$ *then for all* $m \neq n$, there exists a_m *such that* $(a_m, a_n, \mathbf{0}) \sim_f \mathbf{b}$.

disagreement point, which corresponds to DM imagining that players have to find an agreement (infinite cost of disagreement).

The axiom states that it is always possible to equate the fairness of an allocation with payoff to only one individual to that of an initially fairer allocation by giving appropriate payoffs to any second individual. This property requires the fairness ranking never to be satiated in any individual payoff.

Definition: The pair of possible payoffs to individuals m and n is *Preferentially Independent with respect to its Complement* (P.I.C.), if the fairness ranking in the (a_m, a_n)-space is independent of $\mathbf{a}_{-m,n}$.

F_4^N (**Pairwise Preferential Independence**) *For all $m, n \in \{1, .., N\}$, the pair of possible payoffs to individuals m and n is P.I.C.*

Similarly to F_4**,** this axiom must hold if the contribution of one person's marginal private payoff to the fairness of an allocation cannot depend on another person's private payoff level.

Theorem 4 *Assume $N \geq 3$ and that DM is susceptible to shame.*
(i) $\succ$ *and* $\succ_f$ *satisfy* $P_1 - P_5$ *and* F_1, F_2 and F_3^N *respectively, if and only if there exist continuous and strictly increasing functions* $u : \mathbb{R}_+ \to \mathbb{R}$ *and* $\varphi : \mathbb{R}_+^N \to \mathbb{R}$ *such that the function* $U : K \to \mathbb{R}$ *defined as* $U(A) = \max_{\mathbf{a} \in A} [u(a_1) + \varphi(a_1, a_2, ..., a_n)] - \max_{\mathbf{b} \in A} [\varphi(b_1, b_2, ..., b_n)]$ *represents* $\succ$ *and* φ *represents* $\succ_f$.
(ii) $\succ_f$ also satisfies F_4^N *if and only if there exist continuous and strictly increasing functions* $v_1, .., v_N : \mathbb{R}_+ \to \mathbb{R}_{++}$, *where* $v_1, .., v_N$ *are unbounded such that* $\varphi(\mathbf{a}) = \prod_{i1}^{N} v_i(a_i)$.

Theorem 4 is analogous to Theorem 2. For the proof, note that the analogue of Theorem 1 can be established by substituting $\mathbf{a}_{-1}$ for a_2 in the theorem and in the proof, where now $\varphi : \mathbb{R}_+^N \to \mathbb{R}$. To establish the analogue of Theorem 3, namely that there are N increasing unbounded functions $v_1, .., v_N$, such that the fairness ranking $\succ_f$ can be represented by $\varphi(\mathbf{a}) = \prod_{i1}^{N} v_i(a_i)$ if and only if it satisfies F_1, F_2, F_3^N and F_4^N, we first state a stronger version of F_3^N:

$F_3^{N\prime}$ (**Solvability**) *If* $(a_n, \mathbf{a}_{-n}) \not\succ_f$ $\mathbf{b}$ *then for all* $m \neq n$, *there exists* a_m *such that* $(a_m, a_n, \mathbf{a}_{-m,n}) \sim_f$ $\mathbf{b}$.

We observe that *Continuity,* F_1, F_2 and F_3^N imply *Solvability.* To see this, assume

$(a_n, \mathbf{a}_{-n}) \not\succ_f b$. By F_2, $(a_n, \mathbf{0}) \not\succ_f (a_n, \mathbf{a}_{-n})$ and hence (using F_1) $(a_n, \mathbf{0}) \not\succ_f b$. By F_3^N, there exists $\widetilde{a}_m$ such that $(\widetilde{a}_m, a_n, \mathbf{0}) \sim_f \mathbf{b}$. By F_2 again, $(\widetilde{a}_m, a_n, \mathbf{z}) \succeq_f \mathbf{b}$ for all $\mathbf{z} \in \mathbb{R}_+^{N-2}$. Therefore, by *Continuity*, there must be $a_m \in \mathbb{R}_+$ for which $(a_m, a_n, \mathbf{a}_{-m,n}) \sim_f \mathbf{b}$. We can then apply:

Theorem (Luce and Tukey [1964]) *Pairwise Preferential Independence and Solvability imply the existence of an additive representation of* $\succ_f$.

The proof of this theorem can be found in Kranz et al (1971). We illustrate the idea for the case $N = 3$ by showing that F_4^N implies F_4 for (without loss of generality) the pair of individuals 1 and 2, independent of the payoff to individual 3:

For any (a_1^0, a_2^0, a_3^0) and any a_1^1, define a_2^1 and a_3^1 such that

$$\left(a_1^1, a_2^0, a_3^0\right) \sim_f \left(a_1^0, a_2^1, a_3^0\right) \sim_f \left(a_1^0, a_2^0, a_3^1\right).$$

Applying F_4^N twice implies that

$$\left(a_1^1, a_2^1, a_3^0\right) \sim_f \left(a_1^1, a_2^0, a_3^1\right) \sim_f \left(a_1^0, a_2^1, a_3^1\right).$$

For any a_1^2, define a_2^2 and a_3^2 such that

$$\left(a_1^2, a_2^0, a_3^0\right) \sim_f \left(a_1^0, a_2^2, a_3^0\right) \sim_f \left(a_1^0, a_2^0, a_3^2\right) \sim_f \left(a_1^1, a_2^1, a_3^0\right).$$

We have to show that $(a_1^2, a_2^1, a_3) \sim_f (a_1^1, a_2^2, a_3)$ for any value of a_3: $(a_1^2, a_2^0, a_3^0) \sim_f (a_1^1, a_2^0, a_3^1)$, so by F_4^N also $(a_1^2, a_2^1, a_3^0) \sim_f (a_1^1, a_2^1, a_3^1)$. Similarly $(a_1^0, a_2^2, a_3^0) \sim_f (a_1^0, a_2^1, a_3^1)$, so by F_4^N also $(a_1^1, a_2^2, a_3^0) \sim_f (a_1^1, a_2^1, a_3^1)$. Using transitivity, $(a_1^2, a_2^1, a_3^0) \sim_f (a_1^1, a_2^2, a_3^0)$ and by F_4^N this is independent of a_3^0. Hence $(a_1^2, a_2^1, a_3) \sim_f (a_1^1, a_2^2, a_3)$ for any value of a_3.

The existence of utility functions according to which $\succ_f$ is represented by the Nash product follows, as before, where additivity is implied by Luce and Tukey's theorem. We gave the intuition for the remainder of the proof of Theorem 4 after stating Theorem 2.

5.2. A Game of Trust

Consider the game of trust, which is depicted in Figure 2 and is a variant of a game suggested by Tadelis (2008): In the first stage, player 1 can either trust (T) or not trust (N). Action N ends the game and leads to payoff n for both player 1 and 2. Write this outcome as (n, n). If trusted, player 2 can either cooperate (C) or defect (D). Action D generates the outcome $(0, d)$ with certainty. Action C leads to the cooperative outcome $\left(\frac{c}{p}, \frac{c-(1-p)d}{p}\right)$ only with probability p, and to the uncooperative outcome $(0, d)$ otherwise, where $d > c > n > 0$.

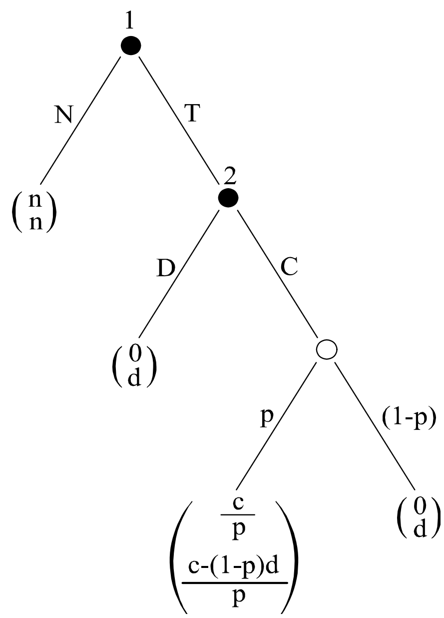

Figure 2: A game of trust with uncertain outcomes.

Player 1 is not susceptible to shame, but player 2 is. Suppose both players are risk neutral and satisfy the expected utility axioms, so that lotteries are evaluated by their expected value. In this case, player 1's options can be written as the two menus $N = \{(n, n)\}$ and $T = \{(c, c), (0, d)\}$.

Unlike the setting considered so far, in which DM chooses an allocation in two stages, the game of trust is a strategic situation: player 1 chooses a menu from which player 2 will choose in the second stage. If instead player 2 chooses over menus in an unobserved first stage, we assume that he evaluates menus according to a variant of our representation,

$$U(A) = \max_{\mathbf{a} \in A} [a_2 + \beta \widetilde{a}_1 \widetilde{a}_2] - \beta \max_{\mathbf{b} \in A} [b_1 b_2]$$

where $\widetilde{\mathbf{a}}$ is player 1's expectation of the allocation $\mathbf{a}$ generated by player 2's choice. As we point out in the discussion of Theorem 2, this suggests that player 2's choice from menu A is governed by maximizing the term $a_1 + \beta \widetilde{a}_1 \widetilde{a}_2$. Player 1, on the other hand, evaluates allocation $\mathbf{a}$ according to her payoff, a_1.

We consider two cases: In the observed case, player 1 observes player 2's action whereas in the unobserved case, player 1 only observes the outcome of the game. In what follows we restrict the strategy space of each player to pure strategies. The next proposition characterizes the (pure strategies) equilibria of this game.[24]

Proposition:

i) In the observed case, the unique equilibrium is T, C, if $\frac{d-c}{\beta c^2} < 1$. If $\frac{d-c}{\beta c^2} > 1$ it is N, D, and if $\frac{d-c}{\beta c^2} = 1$ both equilibria exist.

ii) In the unobserved case, there are no equilibria if $\frac{d-c}{\beta c^2} < p$. If $\frac{d-c}{\beta c^2} \geq p$ the unique equilibrium is N, D.

Proof:

i) In the observed case, trusted player 2 can either choose C, which carries no shame and generates direct utility c, or he can choose D, which carries shame βc^2 and generates direct utility d. Hence player 2 cooperates if $c > d - \beta c^2$, is indifferent between cooperating and defecting if $c = d - \beta c^2$ and defects if $c < d - \beta c^2$. Anticipating this, player 1 chooses T if $c > d - \beta c^2$ and N if $c < d - \beta c^2$. If $c = d - \beta c^2$, there are two equilibria, T, C and N, D.

ii) Consider the two possible equilibria of the unobserved case. Suppose player 2 was required to play C in equilibrium. Then, player 1 expects to see either the outcome (c, c) or $(0, d)$, so neither outcome makes her think that player 2 deviated. Since the expected outcome $\widetilde{\mathbf{a}} = (c, c)$ is not affected by player 2's action, it is profitable for player 2 to deviate and play D, generating a higher direct utility (d instead of c) without increasing shame. Therefore, there is no equilibrium where player 2 chooses C. Suppose then that player 2 is required to play D in equilibrium. In that case, player 1 expects to see the outcome $(0, d)$ for sure, and $\widetilde{\mathbf{a}} = (0, d)$. Playing D, therefore, generates direct utility d and carries shame βc^2. If, however, player 1 observes (c, c), then this can only be explained by player 2 having deviated from D to C,[25] and accordingly $\widetilde{\mathbf{a}} = (c, c)$.[26] If player 2 plays C he receives direct utility c and with probability p there is no shame, while with probability $(1 - p)$ shame is still βc^2. Hence, player 2 is willing to play D in equilibrium, if and only if $d - \beta c^2 \geq c - (1 - p)\, \beta c^2$ or $\frac{d-c}{\beta c^2} \geq p$. In that case player 1 anticipates player 2 to play D and chooses N.■

The interesting case is where $\frac{d-c}{\beta c^2} \in [p, 1)$. Player 1's equilibrium behavior is to trust

[24] It follows immediately from the arguments given in the proof of the proposition that there are no mixed strategy equilibria for this game. Restricting the strategy spaces to pure strategies only serves the purpose of determining out-of-equibrium beliefs, as is needed for part (ii) of the proposition and further explained in the next footnote.

[25] This is true only because we exclude mixed strategies from the players' strategy spaces. Otherwise, (c, c) could be explained by a continuum of mixed strategies and we would have to specify out-of-equilibrium beliefs.

[26] More precisely, $\widetilde{\mathbf{a}}$ is player 2's belief about player 1's perception of his action. Whether or not player 1 actually updates her beliefs in this manner is irrelevant.

(T), if and only if he can observe player 2's action. This can be explained by player 1's correct anticipation of shame as a motivating force, leading player 2 to cooperate *only* under observation. Emotions like altruism or guilt, which do not depend on observation, can not explain this behavior.

6. Related Literature

Other-regarding preferences have been considered extensively in economic literature. In particular, inequality aversion as studied by Fehr and Schmidt (1999) is based on an objective function with a similar structure to the representation of second-stage choice in Theorem 3.[27] Both works attach a cost to any deviation from choosing the fairest alternative. In Fehr and Schmidt's work, the fairest allocation need not be feasible and is independent of the choice situation. In our work, the fairest allocation is always a feasible choice and it is identified through the axioms. This dependence of the fairest allocation on the choice situation allows us to distinguish observed from unobserved choice.

The idea that there may be a discrepancy between DM's preference to behave "pro-socially" and her desire to be viewed as behaving pro-socially is not new to economic literature. For a model thereof, see Benabou and Tirole (2006).

Neilson's (2006-b) work is motivated by the same experimental evidence as ours. He also considers menus of allocations as objects of choice. Neilson does not axiomatize a representation result, but points out how choices among menus should relate to choices from menus, if shame were the relevant motive. He relates the two aspects of shame that also underlie the *Set Betweenness* property in our work; DM might prefer a smaller menu over a larger menu either because avoiding shame compels her to be generous when choosing from the larger menu, or because being selfish when choosing from the larger menu bears the cost of shame.

The structure of our representation resembles the representation of preferences with self-control under temptation, as axiomatized in GP. GP study preferences over sets of lotteries and show that their axioms lead to a representation of the following form:

$$U^{GP}(A) = \max_{a \in A} \left\{ u^{GP}(a) + v^{GP}(a) \right\} - \max_{b \in A} \left\{ v^{GP}(b) \right\}$$

with u^{GP} and v^{GP} both linear in the probabilities and where A is now a set of lotteries. In their context, u^{GP} represents the "commitment"- and v^{GP} the "temptation"-ranking. While the two works yield representations with a similar structure, their domains - and therefore

[27] Neilson (2006-a) axiomatizes a reference-dependent preference, that can be interpreted in terms of Fehr and Schmidt's objective function.

the axioms - are different. In particular, the objects in GP's work are sets of lotteries. They impose the independence axiom and indifference to the timing of the resolution of uncertainty. This allows them to identify the representation above that consists of two functions that are linear in the probabilities. Each of these functions is an expected utility functional. The objects in our work, in contrast, are sets of allocations and there is no uncertainty. The natural way to introduce uncertainty to our model is to treat our representation as the utility function, which should be used to calculate the expected utility of lotteries over sets. Therefore, DM would typically not be indifferent to the timing of the resolution of uncertainty.[28] However, one of GP's axioms is the *Set Betweenness* axiom, $A \succcurlyeq B \Rightarrow A \succcurlyeq A \cup B \succcurlyeq B$. We show that our axioms *Strong Left Betweenness* (P_2), *Shame* (P_3) and *Fairness Ranking* (F_1) imply *Set Betweenness*. Hence, GP's Lemma 2 can be employed, allowing us to confine attention to sets with only two elements.

Our model is positive in nature, but it is interesting to contrast moral or normative elements in its interpretation with those in the context of the temptation literature: In a work related to GP, Dekel, Lipman and Rustichini (2005) write: "*...by 'temptation' we mean that the agent has some view of what is normatively correct, what she should do, but has other, conflicting desires which must be reconciled with the normative view in some fashion.*" According to this interpretation, the commitment ranking is given a normative value. In our work, shame is based on deviating from some fairness norm that tells DM what she should do. This norm conflicts with DM's selfish commitment ranking.

Empirically, the assumption that only two elements of a choice set matter for the magnitude of shame (the fairest available alternative and the chosen alternative) is clearly simplifying: Oberholzer-Gee and Eichenberger (2004) observe that dictators choose to make much smaller transfers when their choice set includes an unattractive lottery. In other words, the availability of an unattractive allocation seems to lessen the incentive to share.

Lastly, it is necessary to qualify our leading example: The experimental evidence on the effect of (anonymous) observation on the level of giving in dictator games is by no means conclusive. Behavior tends to depend crucially on surroundings, like the social proximity of the group of subjects and the phrasing of the instructions, as, for example, Bolton, Katok and Zwick (1994); Burnham (2003); and Haley and Fessler (2005) record. On the one hand, there is more evidence in favor of our interpretation: In a follow-up to the experiment cited in the introduction, Dana et al (2006) verify that dictators do not exit the game if second-stage choice is also unobserved. Similarly, Pillutla and Murningham (1995) find evidence that people's giving behavior under anonymity depends on the information given to the observing recipient. In experiments related to our leading example, Lazear, Malmendier and

[28] In section 5.2 we account for uncertainty, which can be translated into uncertainty over sets.

Weber (2005) as well as Broberg, Ellingsen and Johannesson (2008) even predict and find that the most generous dictators are keenest to avoid an environment where they could share with an observing recipient.[29] Broberg et al further elicit the price subjects are willing to pay in order to exit the dictator game, finding that the mean exit reservation price equals 82% of the dictator game endowment. Further, for the game of trust that inspired the one we consider in section 5.2., Tadelis (2008) experimentally verifies a probabilistic version of our prediction: when moving from a game with no observation to a game with observation, the likelihood of cooperation by player 2 increases and the likelihood of trust by player 1 increases. On the other hand, our interpretation is in contrast to evidence collected by Koch and Normann (2005), who claim that altruistic behavior persists at an almost unchanged level when observability is credibly reduced. Similarly, Johannesson and Persson (2000) find that incomplete anonymity - not observability - is what keeps people from being selfish. Ultimately, experiments aimed at eliciting a norm share the same problem: Since people use different (and potentially contradictory) norms in different contexts, it is unclear whether the laboratory environment triggers a different set of norms than would other situations: Frohlich, Oppenheimer and Moore (2000) point out that money might become a measure of success rather than a direct asset in the competition-like laboratory environment, such that the norm might be "do well" rather than "do not be selfish."[30] More theoretically, Miller (1999) suggests that the phrasing of instructions might determine which norm is invoked. For example, the reason that Koch and Normann do not find an effect of observability might be that their thorough explanation of anonymity induces a change in the regime of norms, in effect telling people "be rational," which might be interpreted as "be selfish." Then being observed might have no effect on people who, under different circumstances, might have been ashamed to be selfish.

7. Appendix

7.1. Proof of Theorem 1

Let $U : K \to \mathbb{R}$ be a continuous function that represents $\succ$. Define $u(a_1) \equiv U(\{(a_1, 0)\})$. By P_1, $u(a_1) = U(\{(a_1, a_2)\})$ independent of a_2, with $u(a_1)$ continuous and strictly increasing.

Let $\varphi : \mathbb{R}^2_+ \to \mathbb{R}$ be a continuous function that represents $\succ_f$.By F_2, φ is also strictly increasing.

[29] This nicely underlines our interpretation of "shame" as a motive.

[30] Surely the opposite is also conceivable: Subjects might be particularly keen to be selfless when the experimentor observes their behavior. This example is just ment to draw attention to the difficulties faced by experimenters in the context of norms.

Claim 1.1 (Right Betweenness): $A \succeq B \Rightarrow A \cup B \succeq B$.

Proof: There are two cases to consider:

Case 1) $\forall a \in A,\ \exists b \in B$ such that $b \succ_f a$. Let $A = \left\{\mathbf{a}^1, \mathbf{a}^2 ..., \mathbf{a}^N\right\}$ and $C_0 = B$. Define $C_n = C_{n-1} \cup \{\mathbf{a}^n\}$ for $n = 1, 2, .., N$. According to F_1, there exists $b \in B$ such that $a^n \not\succ_f b$. By P_3, $C_{n-1} \not\succ C_n$. By negative transitivity of $\succ$, $C_0 \not\succ C_N$ or $A \cup B \succeq B$.

Case 2) $\exists a \in A$ such that $a \succ_f b,\ \forall b \in B$. Let $B = \left\{\mathbf{b}^1, \mathbf{b}^2 ..., \mathbf{b}^M\right\}$. Define $C_0 = A$ and $C_m = C_{m-1} \cup \{\mathbf{b}^m\}$ for $m = 1, 2, .., M$. By P_3, $\forall C$ such that $a \in C$, $C \not\succ C \cup \{\mathbf{b}^m\}$. Hence, $C_{m-1} \not\succ C_m$. By negative transitivity of $\succ$, $C_0 \not\succ C_M$ or $A \cup B \succeq A \succeq B$, hence $A \cup B \succeq B$.$\|$

Combining Claim 1.1 with P_2 guarantees Set Betweenness (SB): $A \succeq B \Rightarrow A \succeq A \cup B \succeq B$. Having established Set Betweenness, we can apply GP Lemma 2, which states that any set is indifferent to a specific two-element subset of it.

Lemma 1.1 (GP Lemma 2): *If $\succ$ satisfies SB, then for any finite set A, there exist $\mathbf{a}, \mathbf{b} \in A$ such that $A \sim \{\mathbf{a}, \mathbf{b}\}$, $(\mathbf{a}, \mathbf{b})$ solves $\underset{\mathbf{a}' \in A}{\max} \underset{\mathbf{b}' \in A}{\min} U(\{\mathbf{a}', \mathbf{b}'\})$ and $(\mathbf{b}, \mathbf{a})$ solves $\underset{\mathbf{b}' \in A}{\min} \underset{\mathbf{a}' \in A}{\max} U(\{\mathbf{a}', \mathbf{b}'\})$.*

Define $f : \mathbb{R}^2_+ \times \mathbb{R}^2_+ \to \mathbb{R}$ such that $f(\mathbf{a}, \mathbf{b}) = u(a_1) - \widetilde{U}(\mathbf{a}, \mathbf{b})$, where $\widetilde{U} : \mathbb{R}^2_+ \times \mathbb{R}^2_+ \to \mathbb{R}$ is a function satisfying:

$$U(\{\mathbf{a}, \mathbf{b}\}) = \max_{\mathbf{a}' \in \{\mathbf{a}, \mathbf{b}\}} \min_{\mathbf{b}' \in \{\mathbf{a}, \mathbf{b}\}} \widetilde{U}(\mathbf{a}', \mathbf{b}') = \min_{\mathbf{b}' \in \{\mathbf{a}, \mathbf{b}\}} \max_{\mathbf{a}' \in \{\mathbf{a}, \mathbf{b}\}} \widetilde{U}(\mathbf{a}', \mathbf{b}').$$ [31]

By definition we have $f(\mathbf{a}, \mathbf{a}) = 0$ for every $\mathbf{a} \in \mathbb{R}^2_+$. Note as well that

$$\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \Rightarrow f(\mathbf{a}, \mathbf{b}) > 0,$$

as otherwise we would have:

$$U(\{\mathbf{a}, \mathbf{b}\}) = \max \left\{ \begin{matrix} u(a_1) - \max\left\{ {f(\mathbf{a},\mathbf{a})=0 \atop f(\mathbf{a},\mathbf{b})} \right\} \\ u(b_1) - \max\left\{ {f(\mathbf{b},\mathbf{a}) \atop f(\mathbf{b},\mathbf{b})=0} \right\} \end{matrix} \right\} \geq u(a_1) - \max \left\{ \begin{matrix} f(\mathbf{a}, \mathbf{a}) = 0 \\ f(\mathbf{a}, \mathbf{b}) \end{matrix} \right\} = U(\{\mathbf{a}\}).$$

For a decision maker who is not susceptible to shame, $U(\{\mathbf{a}, \mathbf{b}\}) = \max\{u(a_1), u(b_1)\}$. Hence setting $f(\mathbf{a}, \mathbf{b}) \equiv 0$ is consistent with her preferences. The following claims help us to further identify f for a decision maker who is susceptible to shame.

---

[31] Note that $\underset{\mathbf{a} \in A}{\max} \underset{\mathbf{b} \in A}{\min} U(\{\mathbf{a}, \mathbf{b}\}) = \underset{\mathbf{a} \in A}{\max} \underset{\mathbf{b} \in A}{\min} \left[\max_{\mathbf{a}' \in \{\mathbf{a}, \mathbf{b}\}} \min_{\mathbf{b}' \in \{\mathbf{a}, \mathbf{b}\}} \widetilde{U}(\mathbf{a}', \mathbf{b}') \right] = \underset{\mathbf{a} \in A}{\max} \underset{\mathbf{b} \in A}{\min} \widetilde{U}(\mathbf{a}, \mathbf{b})$.

Claim 1.2: (i) $[\varphi(\mathbf{a}) < \varphi(\mathbf{b}) \text{ and } a_1 > b_1] \Leftrightarrow \{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$

(ii) $[\varphi(\mathbf{a}) < \varphi(\mathbf{b}) \text{ and } a_1 \leq b_1] \Rightarrow \{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}\}$

(iii) $[\varphi(\mathbf{a}) = \varphi(\mathbf{b}) \text{ and } a_1 > b_1] \Rightarrow \{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$.

Proof: (i) If $\varphi(\mathbf{b}) > \varphi(\mathbf{a})$ then there exists A such that $\mathbf{a} \in A$ and $A \succ A \cup \{\mathbf{b}\}$. As $a_1 > b_1 \Leftrightarrow \{\mathbf{a}\} \succ \{\mathbf{b}\}$, by P_2 $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$. Conversely if $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\}$, then $\mathbf{b} \succ_f \mathbf{a}$ and hence $\varphi(\mathbf{a}) < \varphi(\mathbf{b})$. Further from SB and P_1, $a_1 > b_1$.

(ii) If $a_1 \leq b_1$ then by SB $\{\mathbf{b}\} \succeq \{\mathbf{a}, \mathbf{b}\}$. Since $\varphi(\mathbf{b}) > \varphi(\mathbf{a})$, there is no B such that $\mathbf{b} \in B$ and $B \succ B \cup \{\mathbf{a}\}$, hence $\{\mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}\}$.

(iii) By P_1 $\{\mathbf{a}\} \succ \{\mathbf{b}\}$ and SB $\{\mathbf{a}\} \succeq \{\mathbf{a}, \mathbf{b}\}$. As $\varphi(\mathbf{a}) = \varphi(\mathbf{b})$, using (i) we have $\{\mathbf{a}\} \sim \{\mathbf{a}, \mathbf{b}\}$.$\|$

Let $(\mathbf{a}^*(A), \mathbf{b}^*(A))$ be the solution of

$$\max_{\mathbf{a}' \in A} \min_{\mathbf{b}' \in A} U(\{\mathbf{a}', \mathbf{b}'\})$$

so $(\mathbf{b}^*(A), \mathbf{a}^*(A))$ solves $\min_{\mathbf{b}' \in A} \max_{\mathbf{a}' \in A} U(\{\mathbf{a}', \mathbf{b}'\})$.

Claim 1.3: There exists $\mathbf{b} \in \arg\max_{\mathbf{a}' \in A} \varphi(\mathbf{a}')$ such that $A \sim \{\mathbf{a}', \mathbf{b}\}$ for some $\mathbf{a}' \in A$ and $\mathbf{b}^*(A) = \mathbf{b}$.

Proof: Assume not, then there exist $\mathbf{a}, \mathbf{c}$ such that $A \sim \{\mathbf{a}, \mathbf{c}\}$, $(\mathbf{a}, \mathbf{c}) = (\mathbf{a}^*(A), \mathbf{b}^*(A))$. Therefore,

$$\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{c}\} \sim \{\mathbf{a}, \mathbf{b}, \mathbf{c}\} \sim A^{32} \forall \mathbf{b} \in \arg\max_{\mathbf{a}' \in A} \varphi(\mathbf{a}')$$

and hence $\mathbf{c} \succ_f \mathbf{b}$, which is a contradiction.$\|$

For the remainder of the proof, let $I_f(\varphi) := \{\mathbf{b}' : \varphi(\mathbf{b}') = \varphi\}$. That is, $I_f(\varphi(\mathbf{b}))$ is the $\sim_f$ equivalence class of $\mathbf{b}$. Define

$$Y(\mathbf{a}, \varphi) = \{\mathbf{b}' \in I_f(\varphi) : \{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{b}'\}\}$$

We make the following four observations:

(1) $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$, $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{c}\}$ and $\mathbf{b} \succ_f \mathbf{c}$ imply $\{\mathbf{a}, \mathbf{c}\} \succeq \{\mathbf{a}, \mathbf{b}\}$.

(2) $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$, $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{c}\} \succ \{\mathbf{c}\}$ and $\mathbf{b} \sim_f \mathbf{c}$ imply $\{\mathbf{a}, \mathbf{c}\} \sim \{\mathbf{a}, \mathbf{b}\}$.

[32] Note that if $(\mathbf{a}, \mathbf{c})$ ($(\mathbf{c}, \mathbf{a})$) solves the maximin- (minimax-) problem over A, it clearly solves this problem over the subset $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ for all $\mathbf{b} \in A \backslash \{\mathbf{a}, \mathbf{c}\}$.

(3) $\mathbf{b} \in Y(\mathbf{a}, \varphi)$, $\mathbf{b}' \sim_f \mathbf{b}$ and $\{\mathbf{b}\} \succ \{\mathbf{b}'\}$ imply $\mathbf{b}' \in Y(\mathbf{a}, \varphi)$.

(4) If $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$, $\{\mathbf{b}'\} \succ \{\mathbf{b}\}$ and $\mathbf{b}' \in I_f(\varphi(\mathbf{b}))$, then either $\{\mathbf{a}, \mathbf{b}'\} \sim \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}'\}$ or $\{\mathbf{a}, \mathbf{b}'\} \sim \{\mathbf{b}'\} \succeq \{\mathbf{a}, \mathbf{b}\}$.

To verify these observations, suppose first that (1) did not hold. Then $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{c}\}$ and $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$, hence by SB $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ and therefore $\mathbf{c} \succ_f \mathbf{b}$, which is a contradiction. If (2) did not hold, we would get a contradiction to $\mathbf{b} \sim_f \mathbf{c}$ immediately. Next suppose that (3) did not hold. Then $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\} \succ \{\mathbf{b}'\} \sim \{\mathbf{a}, \mathbf{b}'\}$. Note that by SB $\{\mathbf{b}\} \succeq \{\mathbf{b}, \mathbf{b}'\}$ and, applying SB again , $\{\mathbf{b}\} \succeq \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$. But then $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$, contradicting $\mathbf{b}' \sim_f \mathbf{b}$. To verify (4), assume $\{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{b}'\}$. Then by Claim 1.2 (i) $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}'\} \succ \{\mathbf{b}'\}$ and then by observation (2) $\{\mathbf{a}, \mathbf{b}'\} \sim \{\mathbf{a}, \mathbf{b}\}$. If on the other hand $\{\mathbf{a}, \mathbf{b}'\} \sim \{\mathbf{b}'\}$, then if $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}'\}$, $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{b}\}$ and SB imply $\{\mathbf{a}, \mathbf{b}\} \succ \{\mathbf{a}, \mathbf{b}, \mathbf{b}'\}$, a contradiction to $\mathbf{b}' \in I_f(\varphi(\mathbf{b}))$. Note that by Claim 1.3 we cannot have $\{\mathbf{b}'\} \succ \{\mathbf{a}, \mathbf{b}'\}$.$\|$

Next we claim that $\varphi(\mathbf{b})$ is a sufficient statistic for the impact of $\mathbf{b}$ on a two element set.

Claim 1.4: There exits a function $\widetilde{U}$ satisfying the condition specified above such that $\varphi(\mathbf{b}) > \varphi(\mathbf{a})$ implies $f(\mathbf{a}, \mathbf{b}) = g(\mathbf{a}, \varphi(\mathbf{b}))$, where $g : \mathbb{R}_+^2 \times \mathbb{R} \to \mathbb{R}$ is weakly increasing in its second argument.

Proof: Such $\widetilde{U}$ exists, if and only if $f(\mathbf{a}, \mathbf{b}) = g(\mathbf{a}, \varphi(\mathbf{b}))$ is consistent with $\succ$. Therefore it is enough to consider the constraints $\succ$ puts on f. Given $\mathbf{a}$ and $\mathbf{b}$, look at all $\mathbf{c}$ such that $\varphi(\mathbf{b}) > \varphi(\mathbf{c})$. We should show that $f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{a}, \mathbf{c})$.

First note that if $\varphi(\mathbf{b}) \geq \varphi(\mathbf{a}) \geq \varphi(\mathbf{c})$, then $f(\mathbf{a}, \mathbf{b}) \geq 0 \geq f(\mathbf{a}, \mathbf{c})$ is consistent with $\succ$. If $\varphi(\mathbf{a}) \geq \varphi(\mathbf{b}) > \varphi(\mathbf{c})$, then $0 \geq f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{a}, \mathbf{c})$ is consistent with $\succ$. If $a_1 = 0$, then $f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{a}, \mathbf{c}) \geq 0$ is consistent with $\succ$. Therefore, confine attention to the case where $a_1 > 0$ and $\varphi(\mathbf{b}) > \varphi(\mathbf{c}) > \varphi(\mathbf{a})$.

By Claim 1.2 (i), F_2 and F_3, there exists $\mathbf{b}' \in I_f(\varphi(\mathbf{b}))$ such that $\{\mathbf{a}\} \succ \{\mathbf{a}, \mathbf{b}'\}$. Thus, there are two cases to consider:

1) $Y(\mathbf{a}, \varphi(\mathbf{b})) \neq \emptyset$.

2) $Y(\mathbf{a}, \varphi(\mathbf{b})) = \emptyset$.

Case 1) Suppose $Y(\mathbf{a}, \varphi(\mathbf{b})) \neq \emptyset$. Define $f(\mathbf{a}, \mathbf{b}) := f(\mathbf{a}, \mathbf{b}')$ for some $\mathbf{b}' \in Y(\mathbf{a}, \varphi(\mathbf{b}))$ (note that by observation (2) $f(\mathbf{a}, \mathbf{b}') = f(\mathbf{a}, \mathbf{b}'')$ $\forall \mathbf{b}', \mathbf{b}'' \in Y(\mathbf{a}, \varphi(\mathbf{b}))$ and using observations (3) and (4), this definition is consistent with $\succ$.) If $Y(\mathbf{a}, \varphi(\mathbf{c})) \neq \emptyset$ then by observation (1) $\{\mathbf{a}, \mathbf{c}\} \succeq \{\mathbf{a}, \mathbf{b}\}$ and hence $f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{a}, \mathbf{c})$. If $Y(\mathbf{a}, \varphi(\mathbf{c})) = \emptyset$ then $\forall \mathbf{c}' \in I_f(\mathbf{c})$, $\{\mathbf{a}, \mathbf{c}'\} \sim \{\mathbf{c}'\}$. By F_2 and continuity of $\succ_f$, there exists $\mathbf{c}' \in I_f(\mathbf{c})$ with $c_1' < b_1'$ for some $\mathbf{b}' \in Y(\mathbf{a}, \varphi(\mathbf{b}))$. Then by Claim 1.1, P_1 and observation (1)

$\{\mathbf{a}\} \succ \{\mathbf{a},\mathbf{c}'\} \succeq \{\mathbf{a},\mathbf{b}\} \succ \{\mathbf{b}\} \succ \{\mathbf{c}\}$, so $\mathbf{c}' \in Y(\mathbf{a},\varphi(\mathbf{c}))$. Contradiction.

Case 2) Suppose $Y(\mathbf{a},\varphi(\mathbf{b})) = \emptyset$. Define $f(\mathbf{a},\mathbf{b}) := u(\mathbf{a}_1) - u(0)$. If $Y(\mathbf{a},\varphi(\mathbf{c})) \neq \emptyset$, then $f(\mathbf{a},\mathbf{c}) < u(a_1) = f(\mathbf{a},\mathbf{b})$. If $Y(\mathbf{a},\varphi(\mathbf{c})) = \emptyset$ then $f(\mathbf{a},\mathbf{c}) = u(a_1) = f(\mathbf{a},\mathbf{b})$.$\|$

Let $S := \{(\mathbf{a},\varphi) : Y(\mathbf{a},\varphi) \neq \emptyset\}$. Note that S is an open set.

Claim 1.5: There is $g(\mathbf{a},\varphi)$, which is continuous.

Proof: If $Y(\mathbf{a},\varphi) \neq \emptyset$, then $g(\mathbf{a},\varphi) = u(a_1) - U(\{\mathbf{a},\mathbf{b}\})$ for some $\mathbf{b} \in Y(\mathbf{a},\varphi)$, and is clearly continuous. If $Y(\mathbf{a},\varphi) = \emptyset$, then $\varphi \leq \varphi(\mathbf{a})$ implies $g(\mathbf{a},\varphi) \leq 0$, while $\varphi > \varphi(\mathbf{a})$ implies $g(\mathbf{a},\varphi) \geq u(a_1) - u(0)$. Define a switch point $(\widehat{\mathbf{a}},\widehat{\varphi})$ to be a boundary point of S such that there exists $\widehat{\mathbf{b}} \in \mathbb{R}^2_+$ with $\varphi\left(\widehat{\mathbf{b}}\right) = \widehat{\varphi}$. For $\widehat{\varphi} = \varphi(\widehat{\mathbf{a}})$ define $g(\widehat{\mathbf{a}},\widehat{\varphi}) := 0$ and for $\widehat{\varphi} > \varphi(\widehat{\mathbf{a}})$ define $g(\widehat{\mathbf{a}},\widehat{\varphi}) := u(\widehat{a}_1) - u(0)$.

Consider a sequence $\{(\mathbf{a}^n,\varphi^n)\} \to (\widehat{\mathbf{a}},\widehat{\varphi})$ in S. Pick a sequence $\{\mathbf{b}^{n\prime}\}$ with $\mathbf{b}^{n\prime} \in Y(\mathbf{a}^n,\varphi^n)\,\forall n$. Define $\{b_1^n\} = \left\{\min\left[\frac{1}{n}, b_1^{n\prime}, \widehat{b}_1\right]\right\}$. Define b_2^n to be a solution to $\varphi(b_1^n,b_2^n) = \varphi^n$. By F_2 and F_3, b_2^n is well defined. Note that by observation (3) $\mathbf{b}^n = (b_1^n,b_2^n) \in Y(\mathbf{a}^n,\varphi^n)$. Lastly, let $\widehat{b}_1^n \equiv b_1^n$ and $\widehat{b}_2^n$ be the solution to $\varphi\left(\widehat{b}_1^n,\widehat{b}_2^n\right) = \widehat{\varphi}$. We have $U(\{\mathbf{a}^n,\mathbf{b}^n\}) = u(a_1^n) - g(\mathbf{a}^n,\varphi^n)$. If in the switch point $\widehat{\varphi} = \varphi(\widehat{\mathbf{a}})$, then $U\left(\left\{\widehat{\mathbf{a}},\widehat{\mathbf{b}}^n\right\}\right) = u(\widehat{a}_1)$. By continuity, $U(\{\mathbf{a}^n,\mathbf{b}^n\}) - U\left(\left\{\widehat{\mathbf{a}},\widehat{\mathbf{b}}^n\right\}\right) \underset{n\to\infty}{\to} 0$, hence

$$\lim_{n\to\infty} g(\mathbf{a}^n,\varphi^n) = \lim_{n\to\infty}\left[u(a_1^n) - u(\widehat{a}_1)\right] = u(\widehat{a}_1) - u(\widehat{a}_1) = 0 = g(\widehat{\mathbf{a}},\widehat{\varphi}).$$

If in the switch point $\widehat{\varphi} > \varphi(\widehat{\mathbf{a}})$, then $U\left(\left\{\widehat{\mathbf{a}},\widehat{\mathbf{b}}^n\right\}\right) = u\left(\widehat{b}_1^n\right) = u(b_1^n)$. By the same continuity argument

$$\lim_{n\to\infty} g(\mathbf{a}^n,\varphi^n) = \lim_{n\to\infty}\left[u(a_1^n) - u(b_1^n)\right] = u(\widehat{a}_1) - u(0) = g(\widehat{\mathbf{a}},\widehat{\varphi}).$$

For $\varphi < \varphi(\mathbf{a})$ let $g(\mathbf{a},\varphi) < 0$. This satisfies the constraint on f. So g can be continuous in both arguments and increasing in φ and such that for any sequence $\{(\mathbf{a}^n,\varphi^n)\}$ in S, with $\{(\mathbf{a}^n,\varphi^n)\} \to (\widehat{\mathbf{a}},\widehat{\varphi})$, we have $\lim_{n\to\infty} g(\mathbf{a}^n,\varphi^n) = 0$.$\|$

That the representation satisfies the axioms is easy to verify. This completes the proof of Theorem 1.[33]■

[33] If F_2 and F_3 were only posed on $\mathbb{R}^2_{++}$ as suggested in section 3, we would have to choose $\widehat{b}_1 > 0$ and $b_1^n > 0$ to use these axioms. This is possible for any switch point other than $(\widehat{\mathbf{a}},\widehat{\varphi}) = (\mathbf{0},\varphi(0))$, for which continuity can be established easily.

7.2. Proof of Theorem 2

Theorem 2 and Theorem 4 (i) are analogous, where Theorem 2 covers the case $N = 2$, while Theorem 4 (i) covers the case $N \geq 3$. We prove Theorem 4 (i) below by first establishing that the analogous version of Theorem 1 holds. From there on the proof of Theorem 2 is identical to the proof of Theorem 4 (i), with a_2 substituted for $\mathbf{a}_{-1}$.

7.3. Proof of Theorem 3

Luce and Tukey [1964] prove the necessity and sufficiency of Solvability (which is implied by Negative Transitivity, Weak Solvability, Pareto and Continuity (apply corollary 1 in the text to the case N=2)) and the Corresponding Trade-offs Condition (the label they use for F_4) to admit an additive representation.[34] To see how a proof works, consider the Lock-Step Procedure,[35] as illustrated by Figure 3:

By F_2, $\succ_f$ indifference curves are downward sloping and continuous. Fix (a_1^0, a_2^0) and $a_2^1 > a_2^0$. Recursively construct a flight of stairs between the indifference curves through (a_1^0, a_2^0) and (a_1^0, a_2^1).

In the direction of increasing a_2 (and hence decreasing a_1) :

a_1^n solves $(a_1^n, a_2^n) \sim_f (a_1^0, a_2^0)$. F_3 guarantees that a solution exists whenever $(0, a_2^n) \preceq_f (a_1^0, a_2^0)$. If $(0, a_2^n) \succ_f (a_1^0, a_2^0)$, the flight of stairs terminates.

a_2^{n+1} solves $\left(a_1^n, a_2^{n+1}\right) \sim_f (a_1^0, a_2^1)$. A solution exists by F_3, as $(a_1^n, 0) \prec_f (a_1^0, a_2^1)$ by F_2.

In the direction of decreasing a_2 (and increasing a_1):

a_1^{-n} solves $\left(a_1^{-n}, a_2^{-n+1}\right) \sim_f (a_1^0, a_2^1)$. A solution exists by F_3, as $\left(0, a_2^{-n+1}\right) \prec_f (a_1^0, a_2^1)$ by F_2.

a_2^{-n} solves $\left(a_1^{-n}, a_2^{-n}\right) \sim_f (a_1^0, a_2^0)$. F_3 guarantees that a solution exists whenever $\left(a_1^{-n}, 0\right) \preceq_f (a_1^0, a_2^0)$. If $\left(a_1^{-n}, 0\right) \succ_f (a_1^0, a_2^0)$, the flight of stairs terminates.

By construction $\left(a_1^{n+1}, a_2^{n+2}\right) \sim_f \left(a_1^n, a_2^{n+1}\right)$ and then by F_4, $\left(a_1^n, a_2^{n+2}\right) \sim_f \left(a_1^{n-1}, a_2^{n+1}\right)$. Thus we have constructed a discrete set of points on another indifference curve from the initial two curves. Repeating this procedure we can fill $\mathbb{R}_+^2$ with countable sets of points on countably many indifference curves.

Now consider a particular indifference curve that lies between two members of this set, as illustrated in Figure 4: Define $\left(a_1^{\frac{1}{2}}, a_2^{\frac{1}{2}}\right)$ implicitly by $\left(a_1^{\frac{1}{2}}, a_2^1\right) \sim_f \left(a_1^0, a_2^{\frac{1}{2}}\right)$ and $\left(a_1^{\frac{1}{2}}, a_2^{\frac{1}{2}}\right) \sim_f (a_1^0, a_2^0)$. Construct a flight of stairs between the indifference curves through $\left(a_1^0, a_2^{\frac{1}{2}}\right)$ and through (a_1^0, a_2^0) as described above. Then we have in direction of decreasing

[34] Their theorem is stated in section 5.1 of the text.

[35] See Keeney and Raiffa (1976).

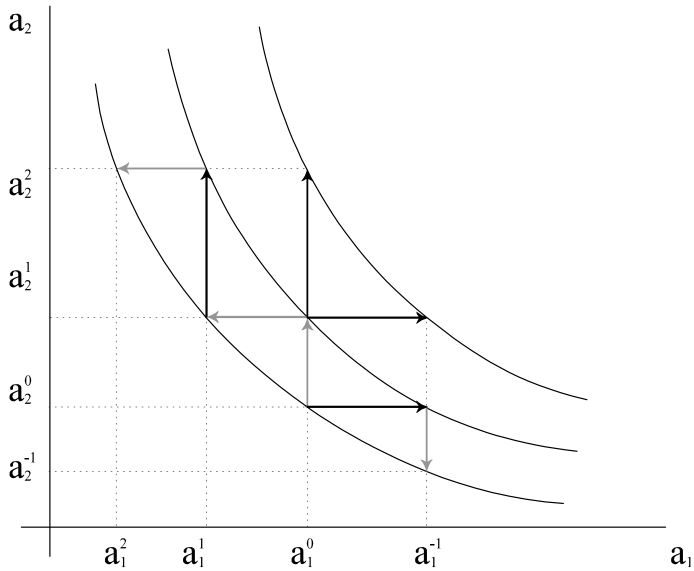

Figure 3: Lock-Step Procedure, Constructing a flight of stairs.

a_2: $\left(a_1^{\frac{n+1}{2}}, a_2^{\frac{n+1}{2}}\right) \sim_f \left(a_1^{\frac{n}{2}}, a_2^{\frac{n}{2}}\right)$ and $\left(a_1^{\frac{n-1}{2}}, a_2^{\frac{n}{2}}\right) \sim_f \left(a_1^{\frac{n+1}{2}}, a_2^{\frac{n+2}{2}}\right)$. Therefore, by construction $\left(a_1^{\frac{n}{2}}, a_2^{\frac{n+1}{2}}\right) \sim_f \left(a_1^{\frac{n+1}{2}}, a_2^{\frac{n+2}{2}}\right)$ and then by F_4, $\left(a_1^{\frac{n-1}{2}}, a_2^{\frac{n+1}{2}}\right) \sim_f \left(a_1^{\frac{n}{2}}, a_2^{\frac{n+2}{2}}\right)$.

Proceed analogously in the direction of increasing a_2.

This demonstrates that if the vertical distance, measured in second component's units, between the indifference curves through (a_1^0, a_2^0) and (a_1^0, a_2^1) in a_1^n is the same as between those through (a_1^0, a_2^1) and (a_1^0, a_2^2) in a_1^{n-1}, then it is also the same between those through (a_1^0, a_2^0) and $\left(a_1^0, a_2^{\frac{1}{2}}\right)$ in $a_1^{\frac{n}{2}}$ and between those through $\left(a_1^0, a_2^{\frac{1}{2}}\right)$ and (a_1^0, a_2^1) in $a_1^{\frac{n-1}{2}}$. Repeating this procedure we can generate a dense set of points on indifference curves that are dense in $\mathbb{R}_+^2$. Then continuity of $\succ_f$ allows us to complete the entire map. Hence, if $(a_1, a_2) \sim_f (a_1', a_2')$ and $(a_1, \widetilde{a}_2) \sim_f (a_1', \widetilde{a}_2')$, then $(\widetilde{a}_1, a_2) \sim_f (\widetilde{a}_1', a_2') \Leftrightarrow (\widetilde{a}_1, \widetilde{a}_2) \sim_f (\widetilde{a}_1', \widetilde{a}_2')$.

As a result, we can create a mapping $a_2 \to \gamma(a_2)$ that transforms the original indifference map to be quasi-linear (vertically parallel indifference curves). The algorithm, which is formally described below, involves proceeding in infinitesimal steps and equalizing the step heights .

Set $\gamma(1) := 0$. To determine $\gamma(a_2)$ for $a_2 > 1$, pick an arbitrary a_1 and let a_1^0 solve $(a_1, a_2) \sim_f (a_1^0, 1 + \Delta)$, where Δ will be infinitesimal for the integration.[36] This solution exists by F_3. Then for every $a_2^* \in (1, a_2]$:[37]
Let a_1^* solve $(a_1^*, a_2^*) \sim_f (a_1^0, 1 + \Delta)$.

[36] As established above, the result of this mapping will be independent of the choice of a_1.
[37] The existence of solutions in the two cases below is guaranteed by the same reasoning as in the above discussion.

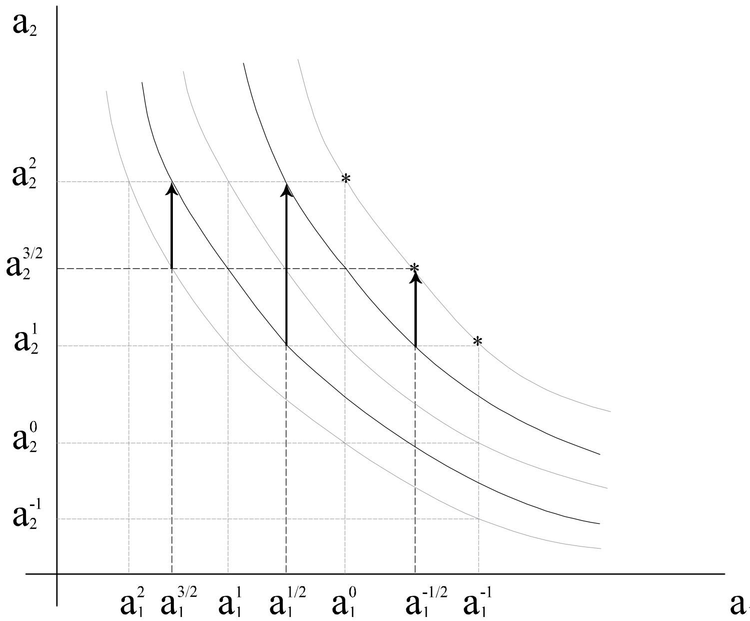

Figure 4: Lock-Step Procedure, Completing the indifference map.

Let a_1^{**} solve $(a_1^{**}, a_2^* + \Delta) \sim_f (a_1^0, 1 + \Delta)$.
Let a_2' solve $(a_1^*, a_2') \sim_f (a_1^0, 1)$.
Let da_2' solve $(a_1^{**}, a_2' + da_2') \sim_f (a_1^0, 1)$.

Note that by F_2, $a_2' < a_2^*$ and $a_2' + da_2' < a_2^* + \Delta$.

Define implicitly $d\gamma(a_2^*) := \widetilde{\gamma}(a_2' + da_2') - \gamma(a_2')$, where

$$\widetilde{\gamma}(a) := \begin{cases} \gamma(a) \text{ for } a \leq a_2^* \\ \gamma(a_2^*) + a - a_2^* \text{ for } a > a_2^* \end{cases}$$

and then

$$\gamma(a_2) := \gamma(1) + \int_1^{a_2} d\gamma(a_2^*) = \int_1^{a_2} d\gamma(a_2^*) \,.$$

Analogously determine $\gamma(a_2)$ for $a_2 < 1$: Pick an arbitrary a_1^0 and let a_1 solve $(a_1, a_2) \sim_f (a_1^0, 1)$. Then for every $a_2^* \in [a_2, 1)$:
Let a_1^* solve $(a_1^*, a_2^*) \sim_f (a_1^0, 1)$.
Let a_1^{**} solve $(a_1^{**}, a_2^* - \Delta) \sim_f (a_1^0, 1)$.
Let a_2' solve $(a_1^*, a_2') \sim_f (a_1^0, 1 + \Delta)$.
Let da_2' solve $(a_1^{**}, a_2' - da_2') \sim_f (a_1^0, 1 + \Delta)$.

Note that $a_2' < a_2^*$ and $a_2' + da_2' < a_2^* + \Delta$ by F_2.

Define implicitly $d\gamma(a_2^*) := \gamma(a_2') - \widetilde{\gamma}(a_2' - da_2')$, where

$$\widetilde{\gamma}(a) := \begin{cases} \gamma(a) \text{ for } a \geq a_2^* \\ \gamma(a_2^*) - a + a_2^* \text{ for } a < a_2^* \end{cases}$$

and

$$\gamma(a_2) := \gamma(1) + \int_1^{a_2} d\gamma(a_2^*) = -\int_{a_2}^{1} d\gamma(a_2^*) < 0.$$

Then $\gamma : \mathbb{R}_+ \to \mathbb{R}$, is a continuous and increasing function. The $\succ_f$ indifference curves are quasi-linear with respect to $\gamma(a_2)$, so there is an increasing continuous function $\xi : \mathbb{R}_+ \to \mathbb{R}$, such that $\xi(a_1) + \gamma(a_2)$ generates the same indifference map. Hence re-defining

$$\varphi(a) := \xi(a_1) + \gamma(a_2)$$

represents $\succ_f$. Define

$$v_1(a_1) := \exp(\xi(a_1)) \text{ and } v_2(a_2) := \exp(\gamma(a_2)).$$

Then $v_1, v_2 : \mathbb{R}_+ \to \mathbb{R}_{++}$ are increasing and continuous and if we re-define, yet again, $\varphi(\mathbf{a}) := v_1(a_1) v_2(a_2)$, it represents $\succ_f$. By F_3, the functions v_1, v_2 must be unbounded.

That the representation satisfies the axioms is easy to verify.■

7.4. Proof of Theorem 4

(i) The analogue of Theorem 1 can be established by substituting $\mathbf{a}_{-1}$ for a_2 in the theorem and in the proof, where now $\varphi : \mathbb{R}_+^N \to \mathbb{R}$.

Let φ be a representation of $\succ_f$. Let $\overline{\varphi} := \sup_{\mathbf{a} \in \mathbb{R}_+^N} \varphi(\mathbf{a})$ and $\underline{\varphi} := \inf_{\mathbf{a} \in \mathbb{R}_+^N} \varphi(\mathbf{a})$, if they are well defined. Otherwise, take $\overline{\varphi} = \infty$ and $\underline{\varphi} = -\infty$.

As before, let $S := \{(\mathbf{a}', \varphi') : Y(\mathbf{a}', \varphi') \neq \emptyset\}$. By F_3^N and the representation analogous to Theorem 1, $u(a_1) - u(0) > g(\mathbf{a}, \varphi)$ for $(\mathbf{a}, \varphi) \in S$.

Let $\succ_S$ be a binary relation on S defined by $(\mathbf{a}, \varphi) \succ_S (\widetilde{\mathbf{a}}, \widetilde{\varphi}) \Leftrightarrow \{\mathbf{a}, \mathbf{b}\} \succ \left\{\widetilde{\mathbf{a}}, \widetilde{\mathbf{b}}\right\}$ $\forall \mathbf{b} \in Y(\mathbf{a}, \varphi)$ and $\forall \widetilde{\mathbf{b}} \in Y\left(\mathbf{a}, \widetilde{\mathbf{b}}\right)$.

Define $U_S : \mathbb{R}_+^N \times (\underline{\varphi}, \overline{\varphi}) \to \mathbb{R}$ such that $U_S(\mathbf{a}, \varphi) := U(\mathbf{a}, \mathbf{b})$ for some $\mathbf{b} \in Y(\mathbf{a}, \varphi)$. By Theorem 1, $\succ_S$ is a weak order that can be represented by U_S. Note that the Consistency axiom (P_5) is relevant precisely on this domain. For $(\mathbf{a}, \varphi) \notin S$ define

$$U_S(\mathbf{a}, \varphi) := \begin{cases} 0 \text{ for } \varphi(\mathbf{a}) < \varphi \\ u(a_1) \text{ for } \varphi(\mathbf{a}) \geq \varphi \end{cases}.$$

Claim 4.1: U_S is continuous in all arguments.

Proof: Since the utility function is continuous on S, and because outside of S the function was chosen to be either a constant (hence continuous) or a continuous function, the only candidates for discontinuity are points on the boundary of S. There are two cases:

Case 1) $\varphi(\mathbf{a}) \geqslant \varphi$: Take $(\mathbf{a}, \varphi) \in bdr(S)$. Since $(\mathbf{a}, \varphi)$ is a boundary point, it must be that $\varphi(\mathbf{a}) = \varphi$. Now let $\{\mathbf{a}^n, \varphi^n\}$ be a sequence in S which converges to $(\mathbf{a}, \varphi)$. By the definition of S, $U_s\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right) = u(a_1^n) - g\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right)$. Because preferences are continuous and using the properties of g from Theorem 1, we have $\lim_{n\to\infty} u(a_1^n) - g\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right) = u(a_1)$ as required.

Case 2) $\varphi(\mathbf{a}) < \varphi$: Take $(\mathbf{a}, \varphi) \in bdr(S)$. Again, let $\{\mathbf{a}^n, \varphi^n\}$ be an arbitrary sequence in S which converges to $(\mathbf{a}, \varphi)$. By the definition of S,

$$U_s\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right) = u(a_1^n) - g\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right) > \inf_{\mathbf{b}} \{u(b_1) : \varphi(\mathbf{b}) = \varphi^n \text{ and } b_1 < a_1^n\}.$$

Since $\succ$ is continuous, we have

$$\begin{aligned} \lim_{n\to\infty} u(a_1^n) - g\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right) &= u(a_1) - g((a_1, \mathbf{a}_{-1}), \varphi) \geq \\ \inf_{\mathbf{b}} \{u(b_1) : \varphi(\mathbf{b}) = \varphi \text{ and } b_1 < a_1\} &= u(0). \end{aligned}$$

where the last equality is implied by F_3^N. As $(\mathbf{a}, \varphi) \notin S$, we claim that $u(a_1) - g((a_1, \mathbf{a}_{-1}), \varphi) \leq \inf_{\mathbf{b}} \{u(b_1) : \varphi(\mathbf{b}) = \varphi \text{ and } \{\mathbf{b}\} \sim \{\mathbf{a}, \mathbf{b}\}\} = u(0)$. If not, then $u(a_1) - g((a_1, \mathbf{a}_{-1}), \varphi) = u(c_1) > u(0)$. But for any $\mathbf{c}$ with $c_1 > 0$, using F_3^N, we could find $\mathbf{c}'$ with $c_1' < c_1$ and $\varphi(\mathbf{c}') = \varphi(\mathbf{c})$. Using Theorem 1, this would imply that $(\mathbf{a}, \varphi) \in S$, which is a contradiction. Combining we have $\lim_{n\to\infty} u(a_1^n) - g\left(\left(a_1^n, \mathbf{a}_{-1}^n\right), \varphi^n\right) = u(0)$, as required.$\|$

Definition: For $(\mathbf{a}, \varphi) \in S$, define $I_S(\mathbf{a}, \varphi) := \{(\mathbf{a}', \varphi') : (\mathbf{a}', \varphi') \sim_S (\mathbf{a}, \varphi)\} \subseteq S$. That is, $I_S(\mathbf{a}, \varphi)$ is the $\succ_S$ equivalence class of $(\mathbf{a},\varphi)$.

Let $a_1^* : \mathbb{R}_+^2 \times (\underline{\varphi}, \overline{\varphi}) \to \mathbb{R}_+$ be the solution to

$$u(a_1^*(\mathbf{a}, \varphi)) = u(a_1) - g(\mathbf{a}, \varphi) = U_S(\mathbf{a}, \varphi).$$

a_1^* is the "first component equivalent" functional on S.[38] Since $u(a_1) > u(a_1) - g(\mathbf{a}, \varphi) >$

[38]Formally, $\forall \mathbf{x} \in \mathbb{R}_+^{N-1}$, $\{(a_1^*(\mathbf{a}, \varphi), \mathbf{x})\} \sim \{\mathbf{a}, \mathbf{b}\}, \forall \mathbf{b} \in Y(\mathbf{a}, \varphi)$

$u(0)$ and $\succ_S$ is continuous, a_1^* is well defined and we have $(\mathbf{a},\varphi) \succ_S (\widetilde{\mathbf{a}},\widetilde{\varphi}) \Leftrightarrow a_1^*(\mathbf{a},\varphi) > a_1^*(\widetilde{\mathbf{a}},\widetilde{\varphi})$.

Claim 4.2: The shame $g(\mathbf{a},\varphi)$ is strictly increasing in φ.

Proof: Assume to the contrary that there is $\varphi' > \varphi$ and $(\mathbf{a},\varphi') \sim_S (\mathbf{a},\varphi)$ for some $\mathbf{a}$. Then for $\varphi' > \varphi'' > \varphi''' > \varphi$ we must have $(\mathbf{a},\varphi'') \sim_S (\mathbf{a},\varphi''')$ as shame is weakly increasing in φ. Now pick $\mathbf{a}'$ such that $(\mathbf{a}',\varphi) \succ_S (\mathbf{a}',\varphi')$ and $(\mathbf{a}',\varphi),(\mathbf{a}',\varphi') \in S$. This is possible by continuity of U_S, since for $\mathbf{a}''$ such that $\varphi(\mathbf{a}'') = \varphi$ the definition of U_S yields $U_S(\mathbf{a}'',\varphi) > U_S(\mathbf{a}'',\varphi')$. Then by P_5, $(\mathbf{a}',\varphi''') \succ_S (\mathbf{a}',\varphi'')$, a contradiction to shame being weakly increasing in φ.‖

Claim 4.3: For all $(\mathbf{a},\varphi)$ and $\widetilde{\varphi} \in (\varphi(a_1,\mathbf{0}),\overline{\varphi})$ there exists $\widetilde{\mathbf{a}}$ such that $(\widetilde{\mathbf{a}},\widetilde{\varphi}) \in I_S(\mathbf{a},\varphi)$.

Proof: Define φ^* implicitly by $U_s((a_1,\mathbf{0}),\varphi^*) = U_s(\mathbf{a},\varphi)$. This is possible by the Intermediate Value Theorem, as $U_s((a_1,\mathbf{0}),\varphi(a_1,\mathbf{0})) = u(a_1) > U_s(\mathbf{a},\varphi) > U_s((a_1,0),\varphi)$, where the last inequality is due to P_4 and Claim 4.2. There are two cases to consider:

Case 1) $\widetilde{\varphi} \geq \varphi^*$: Then $U_s((a_1,\mathbf{0}),\widetilde{\varphi}) \leq U_s(\mathbf{a},\varphi)$ according to the monotonicity of shame. By F_3^N there is $\overline{a}_2(\widetilde{\varphi})$ that solves $\varphi(a_1,\overline{a}_2(\widetilde{\varphi}),\mathbf{0}) = \widetilde{\varphi}$. Then $U_s((a_1,\overline{a}_2(\widetilde{\varphi}),\mathbf{0}),\widetilde{\varphi}) \geq U_s(\mathbf{a},\varphi)$ and by the Intermediate Value Theorem there is $\widetilde{a}_2(\widetilde{\varphi}) \in [0,\overline{a}_2(\widetilde{\varphi}))$ such that

$$U_s((a_1,\widetilde{a}_2(\widetilde{\varphi}),\mathbf{0}),\widetilde{\varphi}) = U_s(\mathbf{a},\varphi).$$

Case 2) $\widetilde{\varphi} < \varphi^*$: Then

$$U_s((a_1^*(\mathbf{a},\varphi),\mathbf{0}),\widetilde{\varphi}) \leq U_s(\mathbf{a},\varphi) \leq U_s((a_1,\mathbf{0}),\widetilde{\varphi}).$$

By the Intermediate Value Theorem there is $\widetilde{a}_1(\widetilde{\varphi}) \in [a_1^*(\mathbf{a},\varphi),a_1]$ such that

$$U_s((\widetilde{a}_1(\widetilde{\varphi}),\mathbf{0}),\widetilde{\varphi}) = U_s(\mathbf{a},\varphi).\|$$

Combining the two cases we see that $\widetilde{\varphi}$ parametrizes a path

$$\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\widetilde{\varphi}) := \begin{cases} (\widetilde{a}_1(\widetilde{\varphi}),\mathbf{0}) & \text{for } \widetilde{\varphi} < \varphi^* \\ (a_1,\widetilde{a}_2(\widetilde{\varphi}),\mathbf{0}) & \text{for } \widetilde{\varphi} \geq \varphi^* \end{cases}$$

of allocations. According to Claim 4.2 $\varphi(\mathbf{a})$ must be strictly increasing along this path. This implies $\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})$ is strictly increasing in its first component for $\widetilde{\varphi} < \varphi^*$ and in its second component for $\widetilde{\varphi} \geq \varphi^*$.

Now we construct a $\succ_S$ indifference class close to the original one:

Claim 4.4: For $\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})$ as defined above, $\widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})$ that solves

$$\left(\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\widetilde{\varphi}), \widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})\right) \in I_S(\mathbf{a}, \varphi + d\varphi)$$

is increasing in $\widetilde{\varphi}$.

Proof: Assume $\widetilde{\varphi}' > \widetilde{\varphi}$. There are two cases to consider:

Case 1) $\widetilde{\varphi}' > \varphi^*$: Then $\widetilde{a}_{1(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right) = a_1$, $\widetilde{a}_{1(\mathbf{a},\varphi)}(\widetilde{\varphi}) \leq a_1$ and $\widetilde{a}_{2(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right) > \widetilde{a}_{2(\mathbf{a},\varphi)}(\widetilde{\varphi})$. P_4 implies

$$\left(\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\widetilde{\varphi}), \widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})\right) \prec_S \left(\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right), \widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})\right).$$

Case 2) $\widetilde{\varphi}' \leq \varphi^*$: Then $\widetilde{a}_{2(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right) = \widetilde{a}_{2(\mathbf{a},\varphi)}(\widetilde{\varphi}) = 0$ and $\widetilde{a}_{1(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right) > \widetilde{a}_{1(\mathbf{a},\varphi)}(\widetilde{\varphi})$. As $\succ_S$ is increasing in a_1,

$$\left(\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\widetilde{\varphi}), \widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})\right) \prec_S \left(\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right), \widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})\right).$$

As shame increases in φ, we must have $\widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}\left(\widetilde{\varphi}'\right) > \widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})$ in both cases.$\|$

Now we define a re-scaling $\varphi \mapsto \gamma(\varphi)$ in order to transform the original indifference map of $U_S(\mathbf{a}, \varphi)$ to be quasi-linear. We proceed similarly to the proof of Theorem 3. Choose $\varphi^0 \in \left(\underline{\varphi}, \overline{\varphi}\right)$ and define $\gamma(\varphi^0) := 1$. Further set $\gamma(\varphi^0 + d\varphi) := 1 + d\gamma$, where $d\varphi$ is infinitesimal. To define $\gamma(\varphi)$ for $\varphi \neq \varphi^0$, pick $\mathbf{a}$ such that $\varphi^*_{(\mathbf{a},\varphi)} < \varphi^0$. As $\varphi^*_{(\mathbf{a},\varphi)} < \varphi$, this implies $\varphi^*_{(\mathbf{a},\varphi)} < \min[\varphi, \varphi^0]$. Choose $\mathbf{a}^0$ such that $(\mathbf{a}^0, \varphi^0) \in I_S(\mathbf{a}, \varphi)$. We will look at the increasing graphs $\widetilde{\varphi}$ and $\widetilde{\varphi + d\varphi}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})$ as defined above. Consider two cases for applying the Lock-Step Procedure:

Case 1) $\varphi > \varphi^0$: Define a climbing flight of stairs between the graphs $\widetilde{\varphi}$ and $\widetilde{\varphi} + d\widetilde{\varphi}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})$ recursively: Let φ^{n+1} solve $\left(\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\varphi^n), \varphi^{n+1}\right) \sim_S (\mathbf{a}^0, \varphi^0 + d\varphi)$. The solution exists by the construction of $\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\varphi^n)$.

Case 2) $\varphi < \varphi^0$: Define a descending flight of stairs between the graphs $\widetilde{\varphi}$ and $\widetilde{\varphi} + d\widetilde{\varphi}_{(\mathbf{a},\varphi)}(\widetilde{\varphi})$ recursively: Let φ^{-n-1} solve $\left(\widetilde{\mathbf{a}}_{(\mathbf{a},\varphi)}(\varphi^{-n-1}), \varphi^{-n}\right) \sim_S (\mathbf{a}^0, \varphi^0 + d\varphi)$.

Then $\gamma(\widetilde{\varphi})$ can be determined analogously to the proof of Theorem 2 by equalizing all step-heights to $d\varphi$ and integrating. Due to P_5 this definition is independent of the choice of $\mathbf{a}^0$.

Now the indifference map of $U_S(\mathbf{a},\varphi)$ is quasi linear in $\gamma(\varphi)$, where $\gamma:\mathbb{R}_{++}\to\mathbb{R}$ is strictly increasing and continuous. Further remember that $U_S(\mathbf{a},\varphi)$ is strictly decreasing in φ. Therefore, there exists $H:\mathbb{R}_+^N\to\mathbb{R}$, such that $H(\mathbf{a})-\gamma(\varphi)$ represents $\succ_S$ on S.

Define $u_S(a_1):=H(\mathbf{a})-\lim\limits_{\varphi\to\varphi(\mathbf{a})}\gamma(\varphi)$. Because of P_1,

$$U(\{\mathbf{a},\mathbf{b}\}):=\begin{cases} u_S(a_1) \text{ if } \{\mathbf{a}\}\sim\{\mathbf{a},\mathbf{b}\}\succ\{\mathbf{b}\} \\ H(\mathbf{a})-\gamma(\varphi(\mathbf{b})) \text{ if } \{\mathbf{a}\}\succ\{\mathbf{a},\mathbf{b}\}\succ\{\mathbf{b}\} \\ u_S(b_1) \text{ if } \{\mathbf{a}\}\succ\{\mathbf{a},\mathbf{b}\}\sim\{\mathbf{b}\}\end{cases}$$

represents $\succ$ confined to the collection of all two element sets. Therefore, $H(\mathbf{a})\equiv u_S(a_1)+\gamma(\varphi(\mathbf{a}))$ must hold. Hence

$$U(A)=\max_{\mathbf{a}\in A}\left[u_s(a_1)+\gamma(\varphi(\mathbf{a}))\right]-\max_{\mathbf{b}\in A}\left[\gamma(\varphi(\mathbf{b}))\right]$$

represents $\succ$ on K, where φ represents $\succ_f$, and u_s and γ are strictly increasing. Since φ represents $\succ_f$, so does $\gamma(\varphi)$. Hence, there is a representation φ of $\succ_f$, such that γ is the identity and

$$U(A)=\max_{\mathbf{a}\in A}\left[u_s(a_1)+\varphi(\mathbf{a})\right]-\max_{\mathbf{b}\in A}\left[\varphi(\mathbf{b})\right]$$

represents $\succ$ on K.

(ii) To establish the analogue of Theorem 3, namely that there are N increasing unbounded functions $v_1,..,v_N$, such that the fairness ranking $\succ_f$ can be represented by $\varphi(\mathbf{a})=v_1(a_1)\cdot\ldots\cdot v_N(a_n)$, if and only if it satisfies F_1,F_2,F_3^N and F_4^N we apply the Theorem of Luce and Tukey, just as in the proof of Theorem 3. It establishes the existence of an additive representation $\xi_1(a_1)+\ldots+\xi_N(a_N)$ of $\succ_f$. Define $v_n(a_n):=\exp(\xi_n(a_n))$ for all $n\in\{1,..,N\}$. Then $v_1,..,v_N:\mathbb{R}_+\to\mathbb{R}_{++}$ are increasing and continuous and if we re-define $\varphi(\mathbf{a}):=v_1(a_1)\cdot\ldots\cdot v_N(a_N)$, it represents $\succ_f$. By F_3^N, the functions $v_1,..,v_N$ must be unbounded.

That the representations satisfy the axioms is easy to verify.■

References

[1] Anderoni, James and John H. Miller (2002) "Giving according to GARP: An experimental test of the consistency of preference for altruism." *Econometrica*, 70, 737-753.

[2] Benabou, Roland and Jean Tirole (2006) "Incentives and Prosocial Behavior." *American Economic Review*, 96(5), 1652-1678.

[3] Bolton, Gary E., Elena Katok and Rami Zwick (1998) "Dictator game giving: Rules of fairness versus acts of kindness." *International Journal of Game Theory,* 27: 269-299.

[4] Broberg, Thomas, Tore Ellingsen and Magnus Johannesson (2007) "Is Generosity Involuntary?" *Economics Letters,* 94, 32-37.

[5] Burnham, Terence C. (2003) "Engineering altruism: A theoretical and experimental investigation of anonymity and gift giving." *Journal of Economic Behavior and Organization,* 50, 133-144.

[6] Buss, Arnold H. (1980) "Self-Consciousness and Social Anxiety." San Francisco, W. H. Freeman.

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

[8] Charnes, Gary and Mathew Rabin (2002) "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics*, 117 (3), 817-870.

[9] Dana, Jason D., Dalian M. Cain and Robin M. Dawes (2006) "What you don't Know Won't Hurt me: Costly (but quiet) Exit in a Dictator Game." *Organizational Behavior and Human Decision Processes*, 100(2), 193-201.

[10] Dana, Jason D., Roberto A. Weber and Jason Xi Kuang (2005) "Exploiting moral wriggle room: Behavior inconsistent with a preference for fair outcomes." mimeo.

[11] Davis, Douglas D. and Charles A. Holt (1993) " Experimental economics." Princeton University Press.

[12] Dekel, Eddie, Barton L. Lipman and Aldo Rustichini (2005) "Temptation Driven Preferences." mimeo

[13] Fehr Ernst, Klaus M. Schimdt (1999) "A Theory of Fairness, Competition and Cooperation." *Quarterly Journal of Economics*, 114, 817-868.

[14] Frohlich, Norman, Joe Oppenheimer and J. Bernard Moore (2001) "Some doubts measuring self-interst using dictator experiments: The cost of anonymity." *Journal of Economic Behavior and Organization,* 46, 271-290.

[15] Gauthier, David and Robert Sugden (editors) (June 1993) "Rationality, Justice and Social Contract: Themes from Morals by Agreement." University of Michigan Press.

[16] Gul, Faruk and Wolfgang Pesendorfer (2005) "The Simple Theory of Temptation and Self-Control." mimeo

[17] —— (2001) "Temptation and Self Control." *Econometrica*, Vol. 69, No. 6, 1403-1435

[18] Haley, Kevin J. and Daniel M.T. Fessler (2005) "Nobody's watching? Subtle cues affect generosity in an anonymous economic game." *Evolution and Human Behavior,* 26, 245-256.

[19] Hammond, Peter J.(1991), "Interpersonal Comparisons of Utility: Why and How They Are and Should Be Made", in Elster and Roemer, "Interpersonal Comparisons of Well Being". Cambridge: Cambridge University Press, pp 200-254.

[20] Johannesson, Magnus and Bjoran Persson (2000) "Non-reciprocal altruism in dictator games." *Economics Letters,* 69, 137-142.

[21] Karni, Edi and Zvi Safra (1998) "The Hexagon Condition and Additive Representation for Two Dimensions: An Algebraic Approach." *Journal of Mathematical Psychology,* 42, 393-399.

[22] Keeney, Ralph L. and Howard Raiffa, with a contribution by Richard F. Meyer (1976) "Decisions with multiple objectives : preferences and value trade-offs". New York : Wiley.

[23] Koch, K. Alexander and Hans-Theo Norman (2005) "Giving in Dictator Games: Regard *for* Others or Regard *by* others?" mimeo.

[24] Krantz, David H., R. Duncan Luce, Patrick C. Suppes and Amos Tversky (1971) " Foundations of Measurments, Vol 1." Academic Press, New York.

[25] Lazear, Edward P., Ulrike Malmendier and Roberto A. Weber (2005) "Sorting in Experiments with Application to Social Preferences." mimeo.

[26] Luce, R. Duncan and John W. Tukey (1964) "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement." *Journal of Mathematical Psychology,* 1,1-27.

[27] Mariotti, Marco (1999) "Fair Bargains: Distributive Justice and Nash Bargaining Theory." *Review of Economic Studies*, Vol.66, 733-41.

[28] Miller, Dale T. (1999) "The Norm of Self-Interest." *American Psychologist*, Vol. 54, No. 12, 1053-1060.

[29] Nash, John F. (1953) "Two-Person Cooperative Games." *Econometrica,* Vol. 21, No. 1, 128-140.

[30] —— (1950) “The Bargaining Problem.” *Econometrica*, Vol. 18, No. 2, 155-162.

[31] Neilson, William S. (2006-a) "Axiomatic Reference Dependence in Behavior towards Others and Toward Risk." *Economic Theory* 28, 681-692.

[32] —— (2006-b) "A Theory of Kindness, Reluctance, and Shame in Dictator Games." mimeo.

[33] Oberholzer-Gee, Felix and Reiner Eichenberger (2004) "Fairness in Extended Dictator Game Experiments". Working paper

[34] Osborne, Martin J. and Ariel Rubinstein (1994) "A course in Game Theory." MIT press, ISBN 0-262-65040-1.

[35] Pillutla, Madan M. and J. Keith Murningham (1995) "Being fair or appearing fair: Strategic behavior in ultimatum bargaining." *Academy of Managment Journal,* 38,1408-1426.

[36] Rawls, John (1971) "A Theory of Justice." *The Belknap Press of Harvard University Press.*

[37] Tadelis, Steve (2008) "The Power of Shame and the Rationality of Trust." mimeo