Moral Costs and Rational Choice:
Theory and Experimental Evidence

James C. Cox[a], John A. List[b], Michael Price[c],
Vjollca Sadiraj[a], and Anya Samek[d]

[a]Georgia State University
[b]University of Chicago, Australian National University, and NBER
[c]University of Alabama, Australian National University, RWI-Leibniz Institute for Economic Research, and NBER
[d]University of California, San Diego and NBER

April 2024

ABSTRACT

The literature exploring other-regarding behavior has uncovered interesting phenomena, yet the extent to which the data are consistent with theory is not well understood. We explain how recent work challenges conventional preference theory and social preferences models but also show that a small part of this research poses a more fundamental challenge of general rational choice theory. We restate classical principles of rational choice (Sen 1971, 1986) to incorporate moral reference points (minimal expected payoffs and endowments) and postulate that choices exhibit monotonicity to the reference points: "moral monotonicity." Data from a new experiment tests rational choice theory and familiar reference point models from prior literature against moral monotonicity theory. We apply our moral monotonicity theory to data from several experiments in previous literature and find general support for the new theory, thus exhibiting the range of fruitful applicability of moral monotonicity theory.

JEL Classifications: C93, D03, D64
Keywords: rational choice, moral reference points, giving, taking, experiment

Corresponding Author: James C. Cox, jccox@gsu.edu

* We thank the John Templeton Foundation under the Science of Philanthropy Initiative and the National Science Foundation (grant number SES-1658743) for research funding and Yongsheng Xu for helpful comments

1. INTRODUCTION

One of the most influential bodies of economics research in the past few decades revolves around whether and to what extent people value fairness, equity, efficiency, and reciprocity. Experimental work has provided evidence that such motivations can be important in creating and determining surplus allocations in markets (see, e.g., Fehr et al., 1993; Bandiera et al., 2005; Landry et al., 2010; Cabrales et al., 2010; Herz and Taubinsky, 2017), with accompanying theoretical models of social preferences providing a framework to model such behaviors (see, e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Andreoni and Miller, 2002; Cox, Friedman and Sadiraj, 2008; Fudenberg and Levine, 2012; Celen et al., 2017; Galperti and Strulovici, 2017).

Within this line of research, pro-social preferences have been elicited using a class of experiments taking the form of dictator games, gift exchange games, public goods games, ultimatum games, and trust games. While such games have shown that social preferences touch many areas of economic interactions, the literature provide much less guidance as to whether individual choices in such settings satisfy deeply held economic tenets.

Our study explores whether a fundamental tenet of rational choice theory – Sen's (1971) contraction property – is satisfied in sharing choices as observed in the economics literature. To understand more deeply the factors that motivate sharing, a number of scholars have augmented the standard dictator game by varying the feasible action set (e.g., List, 2007; Bardsley, 2008; Cappelen et al., 2013; Korenok et al., 2014). These studies report that dictators change their allocations in interesting ways when presented a chance to take as well as to give to others. For example, in the typical dictator game experiment in which "giving nothing" is the least generous act, substantial sums of money are given away (Engel, 2011). Yet, research shows that if subjects are allowed to take as well as give, they give much less to the other player on average.[1]

The first goal of our study is to step back and synthesize what we have learned about the implications for theory from the existing experimental literature on dictator games. We note that the traditional dictator game, wherein more than 60 percent of dictators pass a positive amount of money, does *not* challenge conventional preference theory (Hicks, 1946; Samuelson, 1947; Debreu, 1959) nor revealed preference theory (Afriat, 1967; Varian, 1982; Andreoni and Miller, 2002). We explain that more recent results from this literature (e.g., List, 2007; Bardsley, 2008;

[1] This sentiment is well reflected by Zhang and Ortmann (2014) who report results from a meta-analysis of dictator games that allow a taking option and find, "…an economically and statistically significant negative effect on giving…"

Cappelen et al., 2013) – do challenge conventional preference theory, including social preferences models (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Cox and Sadiraj, 2007)[2] – but they do *not* challenge rational choice theory. In contrast, Korenok et al. (2014) reports an experiment with data that does challenge rational choice theory: the data are inconsistent with a complete and transitive ordering of own and other's payoffs because they are inconsistent with Sen's (1971, 1986) Property α.

Traditional internal consistency conditions, such as Property α, place restrictions on chosen allocations of scarce resources − the consequences of choice − but do not capture ethical constraints on the actions that produce the consequences. We offer morally monotonic choice theory, a modification of the principles of rational choice that is a response to Sen's (1993, p. 495) appeal for extending choice theory beyond an exclusive focus on internal consistency to also incorporate something external to choice behavior such as objectives, values, or norms. In doing so, we construct moral reference points to synthesize insights from prior literature showing the importance of endowments (Korenok et al., 2014) and minimum feasible payoffs (List, 2007; Bardsley, 2008; Cappelen et al., 2013; Krupka and Weber, 2013) on observed allocations.

We report data from an experiment designed to test directly the implications of morally monotonic choice theory – monotonicity in choice with respect to the observable dimensions that define moral reference points. We find support for implications of moral monotonicity that capture observed patterns of sharing; i.e., we find that choices predictably depend upon the elements that define moral reference points.

Our paper contributes to several literatures. First and foremost, our paper extends a body of work exploring the importance of moral costs on sharing and related pro-social behavior. Moral cost models have been suggested in previous work (e.g., Levitt and List, 2007; DellaVigna et al., 2012; Kessler and Leider, 2012; Ferraro and Price, 2013; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2015). Such models, however, incorporate moral costs as directly assumed parameterizations of a "utility function".[3] Our approach differs from this prior work by introducing fundamental principles of "moral choice" that extend conventional rational choice theory. As such, our model provides a more general framework for exploring how moral costs – which reflect the closeness of choices to the most selfish allocations relative to the initial endowments – impact allocations in situations akin to the standard dictator game.

[2] See also experiments by Grossman and Eckel, 2015; Engel, 2011; Zhang and Ortmann, 2014.
[3] We insert quotation marks around "utility function" because the preferences represented may not be transitive.

More broadly, our paper contributes to the "theory speaking to experiment and experiment speaking to theory" research culture that has permeated experimental economics for decades. Consonant with this approach, we advance a theory of moral reference points that is informed by otherwise anomalous data from prior experiments and design an experiment to test the defining property of the new theory – monotonicity of choice with respect to the elements that define moral reference points.

The remainder of our study is structured as follows. Section 2 explores the implications of data from distinct types of dictator games in previous literature for: (a) *homo economicus* preferences theory; (b) other-regarding preferences theory, including (consequentialist) social preferences models; and (c) general rational choice theory. In order to set the stage for our extension of theory, section 3 briefly reviews the defining properties of rational choice: necessary and sufficient conditions for existence of a complete and transitive ordering of choices. Section 4 introduces an extension of the properties of rational choice to incorporate choice monotonicity in moral reference points. Section 5 presents a new experiment designed to test implications of our theory and some prominent alternative theories. Section 6 reports our experimental results and compares their implications for moral monotonicity theory, conventional rational choice theory, and alternative reference-dependent models. Section 7 reports on implications of moral monotoncity for data from related experiments in Korenok et al. (2014), Krupka and Weber (2013), Lazear, Malmendier, and Weber (2012), Oxoby and Spraggon (2008). Section 8 explains the implications of moral monotonicity for some strategic games with contractions: moonlighting and investment games, and carrot/stick, carrot, and stick games. Section 9 concludes.

2. WHAT CAN WE LEARN ABOUT THEORY FROM DICTATOR EXPERIMENTS?

2.1 Experiments in which Behavior is Inconsistent with (Universal) Selfish Preferences

Kahneman et al. (1986) was the first to report a dictator game experiment in the economics literature, giving subjects a hypothetical choice of choosing an even split of $20 ($10 each) with an anonymous subject or an uneven split ($18, $2), favoring themselves. Three-quarters of the subjects opted for the equal split. These findings set in motion three decades of research examining sharing and allocation of surplus in the lab and field. One stylized result that has emerged from the large literature exploring incentivized choices in such settings is that more than 60 percent of subjects pass a positive amount to their anonymous partners and, on average, give more than 25 percent of the total available (Engel, 2011).

Even though some scholars have argued that such giving patterns violate deeply held economic doctrines, it is important to recall that preference order axioms do not uniquely identify the commodity bundles. In a two-commodity case, for example, preferences may be defined over my hotdogs and my hamburgers. Yet, the same formal theory of preferences can be applied to two commodities identified as my hamburgers and your hamburgers or, for that matter, as my monetary payoffs and yours. Identification of the commodities in a bundle is simply an interpretation of preference theory.

In this way, conventional preference theory, either developed as neoclassical preference theory (Hicks, 1946; Samuelson, 1947; Debreu 1959; textbooks) or revealed preference theory (Afriat, 1967; Varian, 1982; textbooks) can be used for agents who are either self-regarding (*homo economicus*) or other-regarding (including preferences for equity). As such, the received results of giving in *standard* dictator games, while inconsistent with *homo economicus*, can be accommodated by conventional preference theory (Andreoni and Miller, 2002; Fisman et al. 2007).

2.2 Experiments in which Behavior is Inconsistent with Convex Preference Theory

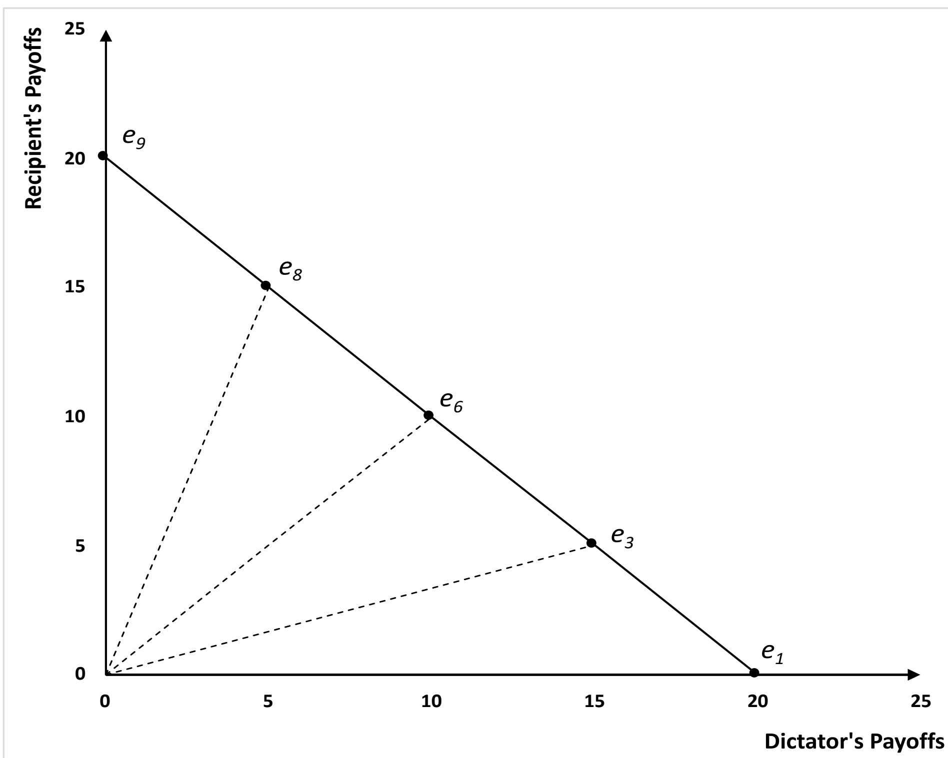
List (2007), Bardsley (2008), and Cappelen et al. (2013), amongst others, use laboratory dictator game experiments to explore how choices are influenced by introducing opportunities for the dictator to take from another subject. Findings from this line of work present a challenge for convex preference theory. For example, List (2007) reports that in his standard Baseline dictator game (which allows sending discrete amounts from \$0 up to \$5), 29% of choices are to send \$0. In List's "Take 1" treatment (standard dictator game augmented to allow taking up to \$1 from the recipient), 65% of the choices are -\$1 (i.e. take 1) or \$0. An implication of conventional preference theory – that includes the strict convexity assumption – is that these (29% and 65%) figures should be the same (because those wanting, but unable to take 1, would choose 0), a pattern that the data clearly refute.

Data from Bardsley (2008), and from the experiment with a representative sample of Danish adult subjects reported by Cappelen et al. (2013) are also clearly inconsistent with convex preference theory including popular models of social preferences. All such models, including inequality aversion (Fehr and Schmidt 1999; Bolton and Ockenfels 2000), CES (Andreoni and Miller 2002), and egocentric altruism (Cox and Sadiraj 2007), have indifference curves that are convex to the origin. Hence, they all have the same implications as conventional convex preference theory for comparisons such as the 29% vs. 65% choices observed in List's experiment. As such, these models are similarly called into question by these dictator games. Convexity of preferences, however, is *not* necessary for choice rationality, so comparisons such as the above for the List, Bardsley, and Cappelen data are uninformative about choice rationality.

2.3 Experiment in which Behavior is Inconsistent with Rational Choice Theory

The experiment reported in Korenok et al. (2014) has implications for (general) rational choice theory, not just special cases with convex preferences. The experimental treatments in Korenok et al. (2014) are illustrated in Figure 1. In all treatments, the feasible set is the same: all admissible (discrete) points on the line extending from 20 on the Dictator's Payoffs axis to 20 on the Recipient's Payoffs axis. The treatments differ in terms of the endowments of the dictator

Figure 1. Endowments for Korenok et al. (2014)



and recipient. In Treatment 1, the endowments are at point e_1 where the dictator is endowed with 20 and the recipient is endowed with 0. In Treatment 9, the endowments are at point e_9 where the dictator is endowed with 0 and the recipient is endowed with 20. In other treatments, the dictator and recipient are endowed with other amounts that sum to 20. The data from choices in Korenok et al. (2014) exhibit a specific pattern with the average recipient payoffs increasing as the endowments e_k move from the horizontal axis towards the vertical axis: \$4.05(e_1), \$5.01(e_3), \$5.61(e_6), \$6.59(e_8), and \$6.31(e_9) with all pairwise differences between these payoffs statistically significant except for the difference between payoffs associated with endowments e_8 and e_9.

Rational choice theory – including the special case of neoclassical and (unconditional) social preferences models − implies that the dictator will make the same choice in all treatments. Such prediction is clearly inconsistent with the data which exhibit systematic differences in allocations across treatments.[4] In order for a model to be consistent with the Korenok et al. (2014) data, it would have to imply that choices of payoffs vary monotonically with endowments. That will be one of the features of the morally monotonic choice theory we offer below. Before developing that new approach, it will be necessary to restate the properties of conventional choice theory that we build on.

3. CONVENTIONAL CHOICE THEORY

Rational choice theory requires that choices satisfy consistency axioms (e.g. Samuelson, 1938; Chernoff, 1954; Arrow, 1959; Sen, 1971, 1986, 1993). We discuss Sen's consistency properties of rational choice.

Notation. Let *X* be the set of all alternatives, and Γ denote the set of all finite subsets of *X*. We call sets, *S* from Γ, "feasible sets" and use *c*(*S*) to denote the choice set; that is the set of all elements chosen from *S*.

Property α (Sen 1971) states that for all feasible sets, $F, G \in \Gamma$

$$\text{Property } \alpha: \text{ If } F \subseteq G \text{ then } \forall\, g^* \in c(G) \cap F \;\Rightarrow\; g^* \in c(F)$$

In words, any element g^* chosen from set G is also chosen from any of its subsets that contain g^*. For singleton choice sets, Property α ("contraction") is equivalent with rationality. For set-value choice functions, one also needs

$$\text{Property } \beta: \text{ If } F \subseteq G \text{ then } \forall x^*, y^* \in c(F),\; x^* \in c(G) \;\Leftrightarrow\; y^* \in c(G)$$

[44] Similar departures from choice theory are documented in a subset of studies, that report endowment effects, that are included in a meta-analysis (Flage (2024) .

In words, Property β ("all or none") states that if some point chosen from *F* is also chosen from a larger set, *G* (containing *F*) then all points chosen from *F* are also chosen from the larger set, *G*. For finite domains, Properties α and β are necessary and sufficient conditions for existence of a complete and transitive order relation: "rational choice theory".

4. MORAL MONOTONICITY CHOICE THEORY

A framework that has been used to describe giving, taking, and related behaviors builds upon the notion of moral cost (Levitt and List, 2007; List, 2007; Lazear et al., 2012; DellaVigna et al., 2012) or concern for norm compliance (Kessler and Leider, 2012; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2015). Using this framework, individuals are said to share with others to avoid experiencing moral cost from failing to do so or from taking actions that are deemed socially inappropriate. We put this approach on a formal foundation that follows from initial work by Cox and Sadiraj (2010).

There are two central features of this approach: (1) definition of moral reference points that are features of the environment (feasible sets, initial allocations); and (2) postulation of principles of choice that are equivalent to traditional principles *when* contractions preserve the moral reference point but differ when they do not.

4.1 Moral Reference Points

In the light of empirical evidence reported above, our definition of moral reference points incorporates two intuitions into the theory of choice: my moral constraints on interacting with you in "the game" we are playing may depend on two features of the environment: (a) the starting position, captured by our endowed (or initial) payoffs (as in Korenok et al. 2014 data);[5] and (b) "surplus" opportunities of the feasible set relative to the payoff each of us can receive at the other's maximal payoff (as in List 2007, Bardsley (2008), and Cappelen et al. 2013 data).[6]

The intuition is straightforward. In making a choice, I may be concerned about whether my choice seems to me (or to others) to be excessively self-serving or (in the traditional definition) "selfish". But which choices are more or less self-serving, or other-serving, cannot be defined without a frame of reference provided by "the game" that you and I are playing.

With respect to feature (a), endowed payoffs are starting points before any choices are

[5] While in many applications the vector of initial endowments is contained in the feasible set of allocations, our approach does not require this restriction.
[6] In this paper, payoff means monetary payoff.

made. If my endowment is 20 and yours is 0 (see Fig. 1) then all actions available to me can be perceived as "altruistic", as they involve giving to you some valuable property that initially belongs to me. In contrast, if my endowment is 0 and yours is 20 (see Fig. 1) then all actions available to me involve taking from you some valuable property that initially belongs to you. As explained above, rational choice theory (including popular social preferences models) makes no distinction between chosen allocations that reflect giving or taking. But data (e.g., Korenok et al. 2014; Flage 2024), anecdotal life experience, and insights from earlier studies (Tversky and Kahneman 1991) suggest that this matters. The question is how to extend rational choice theory to be consistent with such data, experience, and intuition. One way is to introduce monotonicity of choice with respect to ("moral") reference points that depend on endowments, as we do. With this approach, willingness to choose, say, the midpoint of the budget line in Fig. 1 can depend on whether the endowment is on the vertical axis, on the horizontal axis, or at the midpoint itself.

The intuition for feature (b) is also straightforward. If both players have agency, then the outcome will be from the Pareto frontier, such that neither would get less than their payoff at the other's maximum payoff. When the other has no agency, choices of a "morally" concerned dictator would be expected to be from the Pareto set. The closer the choice to my maximal payoff (that is, the closer to other's minimal expected payoff) the more self-serving the choice is, and the lesser the behavior is perceived as "morally" conscious. As an illustrative example, suppose that we are playing the "split \$10" game and I choose payoff of (6, 4), \$6 for me and \$4 for you. How may I feel about that choice? Well, if your *minimum* possible payoff is 0 (corresponding to my maximum possible of \$10), allocating \$4 to you can be perceived as other-serving as it gives you \$4 more than your minimum. Suppose, instead, that there is a constraint whereby the split of \$10 cannot allocate you less than \$4. In this scenario, choice (6, 4) is consistent with self-serving. So, if I want to signal that my choice is motivated by "moral" concerns, (6,4) won't do; I need to go for some other choice that leaves you with more than the minimum feasible amount of \$4.

Data from List (2007), Bardsley (2008), and Cappelen et al. (2013) provide empirical support for the above intuition. In List's experiment, endowments of the dictator and recipient are the same in all treatments. But the expected minimal payoffs differ across treatments. Consider, as an example, the possible choice of dictator payoff of \$10 and recipient payoff of \$5. In List's Baseline treatment, these payoffs correspond to the maximum possible for the dictator and minimum possible for the recipient, the most-self-serving possible choice for the dictator. In contrast, in the Take 5 treatment the choice (10,5) allocates \$5 more to the recipient than their minimum payoff of 0, a seemingly other-serving choice. List's data provide support for the effect of minimum within-game payoffs on chosen payoffs (even when endowments are invariant).

We will formalize these insights and present a concept of moral reference points that depend on *observable* features of the environment: (a) endowments; and (b) minimal payoffs along the (weak) Pareto boundary. Before proceeding, we should note that while the focus of our paper is on dictator games and does not model (experienced or anticipated) reciprocity, our approach has more general applicability for strategic games, as explained in Section 7. Moreover, while the many applications of moral monotonicity in this paper explore settings with two-agent (dictator and strategic, sequential) games, the definition of moral reference points can be extended to *n*-agent environments, as shown in Appendix A.

Let the dictator's opportunity set be some finite set F. The minimal expectation point, P is defined by the minimum payoff when the other gets their maximum feasible payoff in F, that is[7]

$$P_1(F) = \min_{x_1} \max_{x_2} \{x \in F\} \quad \text{and} \quad P_2(F) = \min_{x_2} \max_{x_1} \{x \in F\}$$

Following intuitions (a) and (b), the moral reference point is a function of the initial endowments, e and the minimal expectation payoffs, P. For simplicity, we model it as a weighted average:

$$(*) \qquad f^r = \lambda P + (1-\lambda)e$$

The minimal expectations payoffs and endowments are observable features of the environment. In contrast, the weighting parameter λ may vary across individuals, and is not observable from an experimental design (although it may be estimable from data). As we explain below in section 6.1, the within-subjects feature of our experimental design and our data analysis strategy allow us to test predictions from our theory (stated below) that are *not* dependent on the value of λ.

4.2 Principles of Choice with Moral Reference Points

In this section we extend conventional rational choice theory to incorporate moral reference points.

4.2.a Notation

The domain, Γ of morally monotonic choice functions, $C(\cdot \mid r)$ includes all nonempty finite sets of alternatives. We use notation, $c(S, s^r) \subseteq S$ to denote the non-empty subset of S that choice

[7] This is similar to "minimal expectations payoffs" first used by Roth (1977) in a bargaining framework.

function, $C(\cdot|r)$ assigns to $S \in \Gamma$ in decision problem (S,r), where $r = s^r$ denotes the moral reference point. Superscripts * and ^{r} will be used, respectively, to refer to choices and the reference points. That is, we simply write this decision problem as (S,s^r). Subscripts will be used to denote the coordinate of interest, that is $S_i = \{x_i \mid x \in S\}$ in case of dimension *i*. Ranking, $>$ of scalar sets, *Z* and *Y* in *R*, is defined as follows: $Y > Z$ if $\inf Y \geq \inf Z$ and $\sup Y \geq \sup Z$, that is, the support of *Y* is a right shift of the support of *Z*.

4.2.b Main Principles of Choices

We extend Sen's Properties α and β to incorporate moral reference points and introduce a new (moral monotonicity) property. We postulate that choice sets, $c(S,s^r)$ are *not empty* and satisfy the following three properties for all feasible sets $F, G \in \Gamma$:

Property α_R: If $F \subseteq G$ and $f^r = g^r$, then $g^* \in F \cap c(G,g^r) \Rightarrow g^* \in c(F,f^r)$

Property β_R: If $F \subseteq G$ and $f^r = g^r$, then $\forall x^*, y^* \in c(F,f^r)$, $x^* \in c(G,g^r) \Leftrightarrow y^* \in c(G,g^r)$

Property M_R: For all *i*, if $F = G$ and $f_i^r > g_i^r$, $f_{-i}^r = g_{-i}^r$ then $c_i(F,f^r) > c_i(G,g^r)$

Property α_R ("contraction") and Property β_R ("all or none") are the same as Sen's properties except they are required to hold only for sets and subsets with the same moral reference point. Property M_R ("monotonicity") states that if the moral reference point strictly favors individual *i* then so does the choice set, in the sense that individual *i*'s smallest and largest payoffs are weakly larger.

4.2.c Implications

Appendix B contains detailed discussion of the implications of Properties α_R, β_R, and M_R for choice sets.[8] Here we report two central implications.

We first note that classic results for choice consistency and existence of weak orders can be extended to incorporate moral reference points. We say that a binary relation $\succeq_r$ is constructed from $C(\cdot|r)$ if for all payoff vectors *x*, *y*

[8] As noted above, we do not require the restriction that feasible sets contain initial endowments.

$$x \succeq_r y \text{ if } x \in c(S,r) \text{ for some set } S \in \Gamma \text{ that contains } \{x, y\} \text{ and } r = s^r$$

We say that choice function $C(\cdot \mid r)$ is constructed from a binary relation, $\succeq_r$ if for all $S \in \Gamma$

$$c(S,r) = \{x \in S \mid x \succeq_r y, \text{for all } y \in S \text{ and } r = s^r\}$$

Observation 1 (Weak Orders and Choices). The following statements hold

a. If $C(\cdot \mid r)$ satisfies Properties α_R and β_R then $\succeq_r$ is a weak order.

b. If $\succeq_r$ is a weak order then its choice function, $C(\cdot \mid r)$ satisfies Properties α_R and β_R

Proof: See Appendix B.

We next note that Properties α_R and M_R, together with the assumption that choices are Pareto-efficient, imply that choices are monotonic on moral reference points.

Moral Monotonicity (MM): For all $F \subseteq G \in \Gamma$, and $g^* \in c(G, g^r) \cap F$, for $j \in \{1,2\}$

a. if $f_j^r < g_j^r$ and $f_{-j}^r \geq g_{-j}^r$ then $f_j^* \leq g_j^*$ for some $f^* \in c(F, f^r)$

b. if $f_j^r > g_j^r$ and $f_{-j}^r \leq g_{-j}^r$ then $f_j^* \geq g_j^*$ for some $f^* \in c(F, f^r)$

Proof: See Appendix B.

4.2.d Special Case Representations

The testable inequalities in MM follow from the general Properties α_R, β_R, and M_R. Of course, the testable implications can also be derived from special case models based on the more restrictive assumptions needed for use of a choice function. An example of a choice function that satisfies all three of the above properties is a weighted sum of utilities of payoffs where the weights depend on the moral reference point:

$$c(S, s^r) = \{s^* \in S : U(s^*, s^r) \geq U(z, s^r), \quad \forall z \in S\}$$

where $U(z,r) = \sum_{i=1..n} w_i(r) u(z_i)$ for some increasing function $u(\cdot)$ and weights,

$$\begin{aligned} w_i(r) &= \theta(kr_i) / M(r), \quad \text{if } i \text{ is the decision maker} \\ &= \theta(r_i) / M(r), \qquad \text{otherwise} \end{aligned}$$

for some positive increasing $\theta(\cdot)$, $k \geq 1$ and $M(r) = \theta(kr_i) + \sum_{j \neq i} \theta(r_i)$.[9] If convenient for applications, a researcher can use choice functions consistent with MM.[10]

5. EXPERIMENTAL DESIGN AND PROTOCOL

Our design begins by introducing a symmetric action set in which the dictator can either give to or take from the recipient and compares outcomes in this augmented environment to those observed in dictator games in which the dictator can only give to, or take from, the recipient. The crossed design independently varies the two elements of moral reference points − minimal expectations payoffs and endowments − which allows us to identify their effects on observed choices. We cross this variation in action sets with budget lines having midpoints being mostly below (Inequality), on (Equal) or above (Envy) the 45-degree line. Our approach to identifying the importance of moral reference points shares similarity with Krupka and Weber (2013) who test the importance of norms by comparing final allocations across a standard dictator game and what they call the Bully treatment where the initial endowment is split amongst the dictator and recipient and the dictator is allowed to either give to or take from the recipient.[11]

Across all nine treatments, we restrict the choices of the dictator such that only integer amounts can be given or taken.

5.1 Experimental Design

The experimental design is 3 × 3: (Inequality, Equal, Envy) × (Symmetric, Take, Give). Figure 2 shows three budget lines labeled "Inequality," "Equal," and "Envy." The finite feasible sets in the experiment include discrete points on the lines. Labeling of the feasible sets reflects the location of the midpoints, $B_{j \in \{I,Q,E\}}$ on the lines. The Symmetric treatments have endowment at B_j and permit the dictator to give (move the allocation towards A_j) or take (move the allocation towards C_j). The Take treatments have endowment at B_j and permit the dictator to take (move the allocation towards C_j). The Give treatments have endowment at C_j and permit the dictator

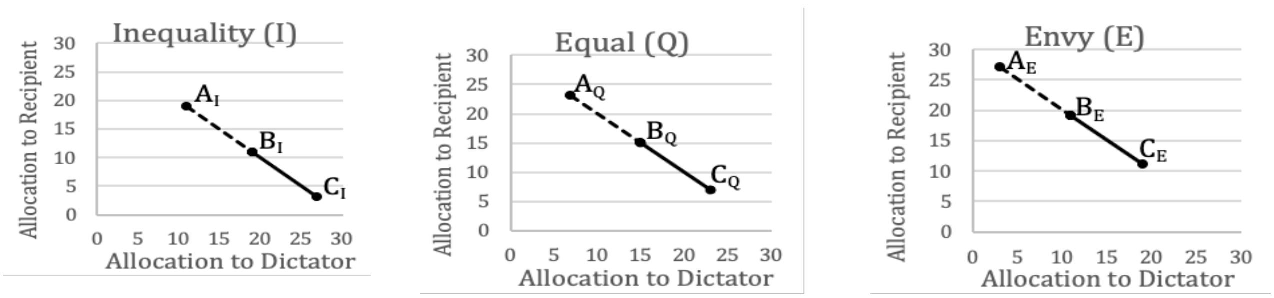
[9] $k > 1$ captures "egocentricity" (Cox and Sadiraj, 2007, 2012), division by M(r) normalizes the weights so they sum to 1.

[10] For illustrative purposes, in Appendix C.1 we use a specification of this choice function to derive testable implications for our dictator game experiment.

[11] Kimbrough and Vostroknutov (2015) use an alternative approach to identify the importance of norms on dictator behavior by eliciting individual-specific measures of norm-sensitivity and correlating this with observed allocations in the standard dictator game.

to give (move the allocation towards B_j). There are two prominent features of this design: (a) the corresponding Take and Give treatments (feasible set $[B_j, C_j]$) have the same minimal expected payoffs but different initial endowments, and (b) a Symmetric treatment's feasible set $[A_j, C_j]$ contains the corresponding Take and Give feasible set $[B_j, C_j]$ as a proper subset (a strict contraction) and has the same initial endowments as Take treatments.

Figure 2. Experiment Feasible Sets: [A, C] for Symmetric and [B, C] for Give or Take



Notes: This figure portrays the feasible allocations for each budget line and action set. Participants in the Give or Take action sets can choose from [B, C], while participants in the Symmetric action set can choose from [A, C]. Actual feasible choices are ordered pairs of integers on the line segments.

The sum of the payoffs of dictator and recipient is the same (30) in all nine treatment cells. In the Inequality-Give treatment (with endowment at point C_I in the left panel of Figure 3): the recipient has an endowment of 3; the dictator has an endowment of 27 and can give up to 8 to the recipient. In the Inequality-Take treatment (with endowment at point B_I in the left panel): the recipient has an endowment of 11; the dictator has an endowment of 19 and can take up to 8 from the recipient. In the Inequality-Symmetric treatment (with endowment at point B_I in the left panel): the recipient has an endowment of 11; the dictator has an endowment of 19 and can give up to 8 or take up to 8. The Equal and Envy treatments change the locations of the (point B or point C) endowments but preserve the Give, Take, or Symmetric action sets. In the Equal feasible set, the Symmetric and Take endowment (at point B_Q in the middle panel) is 15 for the recipient and 15 for the dictator. In the Envy feasible set, the Symmetric and Take endowment (at point B_E in the right panel) is 19 for the recipient and 11 for the dictator.

In the Inequality-Symmetric and Envy-Give treatments, the dictator faces an allocation decision over a budget line that crosses the 45-degree line, as in most standard dictator games. In the Equal-Take and Equal-Symmetric treatments, the initial endowment lies on the 45-degree line.

However, the treatments differ in that the budget line for the Equal-Take treatment lies on and below the 45-degree line whereas the budget line for the Equal-Symmetric treatment crosses the 45-degree line. Data from the three budget lines inform on whether choice response to changes in endowment and minimal payoffs is observed on different parts of the payoff space.

Table 1 shows initial endowments and minimal expectations payoffs that are observable features of the nine treatments in this experiment as well as number of observations and the average payoffs of "dictator" subjects (with and without "selfish" subjects' choices). We observe between 58 to 82 allocations in each of our treatments. About 73% of choices are non-selfish (that is, the dictator did not choose allocation *C*), which is consistent with behavior observed in previous studies (Engel, 2011). In section 6.1, we report tests of choice responses to changes in initial endowments and minimal expectations payoffs.

Table 1. Summary Statistics

Budget Line	Action Set	Initial Endowment	Minimal Expectation Payoff	Nobs	Average Dictator Payoff[a]	
					All Choices	Non-Selfish Choices[b]
Inequality	Give	(27, 3)	(19, 3)	61	22.5 {2.96}	21.6 {2.36}
	Take	(19, 11)	(19, 3)	81	22.8 {3.34}	21.1 {2.30}
	Symmetric	(19, 11)	(11, 3)	82	20.9 {4.95}	19.0 {4.12}
Equal	Give	(23, 7)	(15, 7)	66	20.4 {2.24}	19.4 {1.80}
	Take	(15, 15)	(15, 7)	58	19.8 {2.88}	18.3 {2.22}
	Symmetric	(15, 15)	(7, 7)	62	19.1 {3.52}	17.3 {2.82}
Envy	Give	(19, 11)	(11, 11)	67	16.6 {1.65}	15.9 {1.14}
	Take	(11, 19)	(11, 11)	69	16.9 {1.85}	15.7 {1.15}
	Symmetric	(11, 19)	(3, 11)	66	16.4 {2.55}	15.5 {2.38}

Note. In all treatments, the sum of dictator and recipient payoffs is 30. [a] Standard deviations in brackets; Nobs is the number of observations in each treatment (each subject made only two decisions). [b]selfish choices (allocations C) are not included in this column.

5.2 Protocol for the Experiment

The experiment was conducted in the laboratory of the Experimental Economics Center at Georgia State University using students recruited from the student body at Georgia State. The experiment was approved by the Institutional Review Board of Georgia State University. When they agreed to participate, subjects knew only that they would be in an economics experiment, but not the exact nature of the experiment. Subjects were given as much time as they wanted to read instructions on their computer monitors. After they were finished reading, summary instructions were projected on a screen and read aloud by an experimenter to make clear that all subjects were given the same information about the decision task. All subjects participated in two practice dictator decisions (without payments) to become familiar with both the underlying allocation task and the computer interface. No information was given to subjects about others' practice decisions. After the practice decisions were completed, subjects were informed that the computer would randomly assign them to be active decision makers or passive recipients and that this information would appear on their screen before the start of the first actual round of play.

Subjects were informed that there was no show-up (or non-salient participation) fee in this experiment.[12] Subjects were further informed that each active subject would make two decisions while paired with the same recipient and that one of the two decisions would be randomly selected for payoff. It was explained that these pairings were anonymous and that participants would not know the identity of the person with whom they were paired. A subject made decisions in Give and Take action sets for the same (Equal or Inequality or Envy) budget line; or the subject made decisions in Symmetric and Give or Take action sets for the same budget line. The order of the "games" each active subject faced was independently randomly selected. Subjects were asked to complete a short survey after all decisions were made. Once all subjects had completed the survey, they were paid individually and in private their earnings for the randomly chosen decision round. Subject instructions and the survey are available online: http://excen.gsu.edu/jccox/instructions.

5.3 Participation in the Experiment

In total, we had 612 subjects (306 dictators) participate in the experiment. None of the "dictators" had previous experience (as either dictator or recipient) in dictator games. Each session lasted approximately 50 minutes and each dictator made two decisions. The actual payoffs (from the randomly selected payoff rounds) for dictators were: \$19.46 (average) with the range \$8 (minimum) to \$27 (maximum).[13] Regarding demographic characteristics for our sample, we have 51.6% African-American, 37% males, 43% Social Science study major and 29% STEM, and (26%, 24%, 29%, 19%) reported to be Freshman, Junior, Sophomore and Senior.[14] As noted in Flage (2024), there are differences in the extent of give-take asymmetries

---

[12] The feasible sets for salient payoffs were constructed so that no subject in any treatment could leave the lab with less than \$3.

[13] Recall that the sum of payoffs to dictator and recipient pairs is always \$30.

[14] See Appendix D for more details on the demographics of our decision-maker subjects.

across various demographic dimensions such as gender. Given that our central tests of moral monotonicity rely upon within subject comparisons, we do not need to focus on demographic effects or potential imbalances in observable characteristics across treatments.

5.4 Experimental Design Discriminates Among Consequences, Actions, and Reference Points

Standard models of choice behavior are “consequentialist” models in which only the allocations (or in our experiment, the vector of payoffs) resulting from decisions are important, not the actions which produce those allocations. Of course, our experiments test this common feature of consequentialist models because choosing a specific allocation of payoffs in one treatment requires a different action than in another treatment. For example, allocation (18, 12) is in the interior of all feasible sets in 6 treatments involving Equal and Envy budget lines. In the experiment a dictator can choose this allocation, in which she gets 18 and the recipient gets 12, by giving in Give action set (5 in the Equal-Give or 1 in Envy-Give) or by taking in Take or Symmetric (3 in the Equal-Take and Equal-Symmetric or 7 in Envy-Take and Envy-Symmetric). A consequentialist model will predict that an individual who takes in these Take/Symmetric treatments will give in the above Give treatments because they yield the same allocation of payoffs, so we should see no treatment effect. Yet, our data are at odds with this prediction. The percentage of choices that results in a final allocation of (18, 12) vary from 6.9 to 12.1 across the three action sets in the Equal budget line and from 2.9 to 15.2 percent for the three action sets in the Envy budget line. Compared to the Take action set, the estimated odds ratio of such choice is 2.37 (two-sided p-value = 0.076) for the Give action set and 3.08 (two-sided p-value = 0.018) for the Symmetric action set.

Similarly, an agent who gives 4 in the Inequality-Symmetric treatment is predicted to give 8 in Equal-Give and take 4 in the Envy-Take treatment because all these actions provide the same payoff of 15 for both the dictator and recipient. Again, our data are at odds with such prediction. The percentage of subjects choosing the equal allocation range from 5 percent in the Equal-Give treatment to a high of 30 percent in the Envy-Take treatment. Compared to the Equal-Give treatment, the estimated odds ratio of choosing the equal split is 5.7 (two-sided p-value = 0.008) for the Inequality-Symmetric treatment and 9.5 (two-sided p-value = 0.001) for the Envy-Take treatment. Such differences are at odds with any consequentialist model and imply that behavior in our experiment cannot be rationalized by standard models of choice.

As explicated in the examples, whether choosing a specific allocation requires an action of giving or taking depends on the endowment, and one reason predictions from moral monotonicity differ from those from consequentialist models is the dependence of moral reference points on endowments. There are, of course, well-known models of endowment effects on choice in the extensive literature pioneered by Tversky and Kahneman (1991). Our experimental design, however, discriminates between the effects on choice of endowments and the effects of minimal expectations payoffs, the other element of moral reference points. For example, in each of the three environments (Inequality, Equal, and Envy) in the experiment, the Take and Symmetric treatments have the same endowment (at a point B in a panel of Figure 2), so any endowment effect on choice (from [B,C]) between the two treatments will be the same. In contrast, as explained above, each Take treatment has a different minimal expectations point than the associated Symmetric treatment with the same endowment. Hence, moral monotonicity predicts specific patterns in choices in such Take and Symmetric treatments that are not endowment effects.

6. TESTS OF ALTERNATIVE MODELS

6.1 Tests of the Moral Monotonicity Model vs. Rational Choice

Table 1 reports endowments, minimal expectations payoffs, summary statistics of dictators' payoffs, and number of observations for the nine treatments in our experiment.

6.1.a Within-Subjects Tests

For choices within each budget line, the within subjects responses reflect changes in one of the components of the moral reference point in a direction that is the same for all possible values of the λ weight in statement (*); therefore the prediction from moral *monotonicity* is invariant to individual's value of λ. This feature of our design is evident in Table 1. Consider, for example, the Give, Take and Symmetric rows for the Equal budget line in the table. The change in moral reference point between action sets Give and Take reflects the change in initial endowment from (23,7) to (15,15) with minimal expectations payoffs fixed at (15,7). So the direction of change in moral reference point is the same for any value of λ. Next, the change in moral reference point between action sets T and S reflects the change in minimal expectations payoffs from (15,7) to (7,7) with fixed endowments at (15,15), which implies the same direction of change in moral reference point for any value of λ. A similar explanation (of λ- independence) for predictions applies for all other budget lines in Table 1.

Table 2 reports one-sided *p*-values for the paired *t*-test and *sign* test applied to the decision maker's two choices, separately for subjects who participated in Inequality, Equal and Envy budget lines. The null hypothesis from rational choice theory, including consequentialist social preference models, is no difference in dictator payoffs across the different action sets for any given budget line. The alternative hypothesis from moral monotonicity theory is that dictator payoffs are higher (resp. lower) when the moral reference point becomes more favorable for the dictator (resp. recipient). Hence, tests of moral monotonicity theory center around comparisons of whether dictators earn the most in Give versions of the game and the least in the Symmetric versions of the game.

Table 2. Within-Subjects, Within-Budget Lines Tests

Budget Lines	**Dictator Payoff Means** Moral Reference point is more favorable to the dictator (first entry) or the recipient (second entry)	**One-sided p-val.**		**Nr of Subjects**
		t-test	**sign test**	
Inequality	(23.43, 22.47)	0.005	0.017	94
Equal	(20.08, 19.44)	0.023	0.016	93
Envy	(16.74, 16.51)	0.138	0.442	101

Notes. Each subject participated in only two treatments and made only one payoff decision in each. Choices from 18 (out of 306) subjects who gave in Symmetric and choose B (in Take or Give) are not included as we cannot conclude whether their choice B in Take or Give indicates less generosity (than in Symmetric), or is a consequence of the constraint in Take/Give design.

This pattern is consistent with the dictator's average payoff being about \$1 higher (23.43 – 22.47) when the budget line is Inequality; a difference that is statistically significant at the $p < 0.01$ level using both a matched pairs t-test and the non-parametric sign test. The differences in the Equal and Envy rows of Table 2 – \$0.64 and \$0.23, respectively – are as predicted by moral monotonicity theory. However, the difference is statistically significant for the Equal budget line (p-values < 0.05 using both the matched-pairs t-test and the sign test), but not for the Envy budget line (p-value = 0.138 for the matched pairs t-test and 0.442 for the sign test).

Table 3 reports direct tests of the predicted effects of moral reference points on choices. Conventional rational choice theory predicts all of these effects are 0. There are two observable components of moral reference points, minimal expectations payoffs and endowments. We test for the effect of each component separately as well as jointly on choices. Data from our experiment show statistically significant joint effect of the two dimensions (\$0.77 more for the dictator in Give compared to Symmetric), significant choice monotonicity in minimal expectations payoff ($0.90 more for the dictator in Take than in Symmetric), but the endowment effect ($0.12 in Give vs. Take) is not statistically significant. The tests in the first and third rows of Table 3 are tests of the formal statement of Moral Monotonicity in section 4.2.c. The test in the second row is a direct test of Property M_R in section 4.2.b.

Table 3. Within-Subjects, Between-Budget Lines Tests

Moral Reference Point Components	**Dictator Payoff Means** Moral Reference point is more favorable to the dictator (first entry) or the recipient (second entry)	**One-sided p-val.**		**Nr of Subjects**
		t-test	**sign test**	
Minimal Expectations (Take vs. Symmetric)	(20.53, 19.63)	0.006	0.056	99
Endowments (Give vs. Take)	(19.79, 19.67)	0.315	0.333	96
Minimal Exp. & Endowments (Give vs. Symmetric)	(19.66, 18.89)	0.002	0.008	93

Notes. Each subject participated in only two treatments and made only one payoff decision in each. Choices from 18 subjects who gave in Symmetric and choose B (in Take or Give) are not included as we cannot conclude whether their choice B in Take or Give indicates less generosity (than in Symmetric), or is a consequence of the constraint in Take/Give design.

6.1.b Between-Subjects Tests

The within-subjects tests in Table 3 for effects of changes in minimal expectations payoffs reflect only changes in the *dictators'* minimal expectations payoffs *within* each of the three budget lines (see Table 1). We can test for effects of variation in the *recipient's* minimal expectation payoff using between-subjects, between budget-lines data. Note that the feasible set in the Envy-Give treatment is a subset of the feasible set in the Inequality-Symmetric treatment. Both have the same initial endowments, (19,11) and dictator's minimal expectation payoff, 11 but the recipient's minimal expectation payoff is 11 in Envy-Give and 3 in Inequality-Symmetric. MM predicts that dictators' choice allocates a higher payoff to the recipient in Envy-Give than in Inequality-Symmetric. In terms of the dictator's payoff, the last statement means a lower payoff for the dictator in Envy-Give, as the total amount to be allocated is always $30. Table 4 reports results from a tobit regression using between-subjects data from these two treatments. The dependent variable is dictator's payoff. Tobit regression is used because the experimental design places bounds on possible choices of the dependent variable. The right hand variable is a dummy variable for observations from treatments with the Envy-Give treatment; the reference group is the Inequality-Symmetric treatment. The columns report estimates with or without demographics and with or without a dummy variable for round 2 decisions. Results in all columns support the same conclusions. Consistent with moral monotonicity theory, dictator's payoff is decreasing in recipient's minimal expectations payoff, as indicated by the negative sign of the dummy variable on the Envy-Give treatment.

Table 4. Between-Subjects, Between-Budget Lines Tests

Dep.Var.: Dictator's Payoff	(1)	(2)	(3)
Recipient's Minimal Payoff [-]			
(Envy-Give)	-2.19***	-2.30***	-2.23***
	(0.557)	(0.542)	(0.534)
Constant	19.16***	18.47***	20.73***
(Inequality-Symmetric)	(0.422)	(0.439)	(2.041)
Round 2 (D)	no	yes	yes
Demographics	no	no	yes
Observations	149	149	149
(left, un-, right) censored obs	(2,81,66)	(2,81,66)	(2,81,66)
Log-Likelihood	-257.1	-252.2	-249.3

Notes. One observation per subject; 82 subjects in Inequality-Symmetric and 67 subjects in Envy-Give. Initial endowments are (19,11) in both treatments. Dictator's minimal expectation payoff is 11 in both treatments. Recipient's minimal expectation payoff is 3 in Inequality-Symmetric treatments and 11 in Envy-Give treatment. MM predicted sign in square brackets. Low bound is 11 in both treatments. The largest feasible payoff for the dictator is 19 in Envy-Give and 27 in Inequality-Symmetric. To control for this, the Tobit regression is conducted with the upper bound set at 19 in both treatments. Standard errors in brackets *** p<0.01, ** p<0.05, * p<0.1

6.2 Tests of Some Other Models

We briefly examine implications of alternative models of behavior including models of reference dependence, sharing and sorting, and social image signaling.

Other Reference Dependent Models. The status quo (initial endowment) is the reference point in the classical loss-aversion reference dependent model of Tversky and Kahneman (1991). This TK model predicts that the dictator's final payoff allocation in Give treatments is larger than in the Take treatments, which is the same as our moral monotonicity prediction. This is so because in the Give scenario all feasible allocations introduce loss on dictator's dimension and gain on recipient's dimension whereas in the Take scenario, all feasible allocations offer gain for dictator's payoff but loss for recipient's payoff. However, in the Symmetric and Take scenarios the status quo (the initial endowment) is the same, and therefore the prediction of the Tversky and Kahneman model, for people who do not give in Symmetric, is the same as the conventional rational choice theory prediction. Moral monotonicity predicts a larger final allocation for the dictator in the Take scenario when choice in Symmetric is between B and C in Figure 2. Our data reject the Tversky and Kahneman null hypothesis of no difference in payoffs in favor of the moral monotonicity alternative hypothesis of positive difference in payoffs (see Table 3, first row).

The Koszegi and Rabin (2006) model of reference dependence has recently seen a surge in applied work. Predictions of this model for our games are similar to conventional rational choice theory because, in deterministic settings, optimal "consumption" derived for the conventional preferences model is the "preferred personal equilibrium".[15] Because our data are inconsistent with conventional theory, the data are also inconsistent with the Koszegi and Rabin model.

Sharing and Sorting. Lazear et al. (2012) offer a model of sharing that depends on the environment, where an indicator variable takes value 1 when the environment allows sorting and 0 otherwise. In all of our treatments sorting is not available (i.e., people cannot sort in or out of participating in the games), hence implications of their model for play in our games are similar to standard preference theory, which is inconsistent with our data.

To summarize, our data provide evidence at odds with standard rational choice theory. The data are also at odds with three alternative behavioral models that have often been used to explain sharing. Viewed in its totality, we thus believe our data provide compelling evidence that observable moral reference points matter, and influence choice in a manner consistent with moral monotonicity.

7. IMPLICATIONS OF MORAL MONOTONICITY FOR OTHER TYPES OF DICTATOR GAMES

We have applied moral monotonicity in analysis of data from our experiment. It has broader implications for choice in a range of related experiments including standard (give-only) dictator games and other dictator games that compare the effects of give versus take actions on choices (Korenok et al. 2014), the "bully" dictator game (Krupka and Weber 2013), dictator games with

[15] See Proposition 3 in Koszegi and Rabin (2006, pg.1145).

outside options (Lazear, Malmendier, and Weber 2012), and dictator games where property rights and endowments are earned (Oxoby and Spraggon, 2008; Korenok et al., 2017).

7.1 Give and Take: Moral Monotonicity vs. Warm

Korenok et al. (2014) report a dictator game experiment to test the theoretical model of warm glow developed by Korenok et al. (2013). In particular, the authors explore the effects of changing endowments and framing actions as giving to or taking from the recipient. Korenok et al. (2014) explain that data from their experiment is inconsistent with the predictions of their model. We explained in section 2.3 that their data is inconsistent with conventional rational choice theory. Here we explain that their data is largely consistent with moral monotonicity. In all of their treatments, the minimum expectations point is the natural origin (because the fixed budget line intersects both axes), therefore, changes in moral reference points in their design are entirely determined by changes in endowment. As the endowments, e_j move northwest along the budget line the moral reference points move northwest, favoring the dictator less and less (and the recipient more and more). The average recipient payoffs with the endowments e_j are \$4.05(e_1), \$5.01(e_3), \$5.61(e_6), \$6.59(e_8), and \$6.31(e_9). Moral monotonicity requires dictator's choices to decrease the amount allocated to oneself from scenario 1 to 9, while conventional rational choice theory requires the choices be the same. Korenok, et al. (2014) data reject the implication of conventional theory in favor of moral monotonicity in the three out of four comparisons where differences between treatments are significant (all except the difference between \$6.59 and \$6.31).

7.2 Moral Monotonicity and Bully Games

Moral monotonicity predicts both dictator game choices and social norms elicited by Krupka and Weber (2013). In their experiment, the moral reference point favors the dictator more in the standard dictator game than in the bully dictator game. Hence, moral monotonicity requires choices in the bully treatment to be drawn from a distribution that is less favorable to the dictator than the distribution of choices in the standard game. So, we expect a higher amount allocated to the recipient and a positive estimate of the bully treatment in an ordered logistic regression. The reported mean amounts allocated to the recipients are \$2.46 (standard) and \$3.11 (bully) and the coefficient estimate for the bully treatment is significantly positive in their Table 2.

Moreover, the distribution of elicited norms reported in Krupka and Weber's Table 1 are also consistent with moral monotonicity. A paired *t*-test of the two distributions rejects the null hypothesis of no effect (implied by conventional rational choice theory), in favor of the moral monotonicity-consistent alternative (approval of higher allocations to recipients). Hence, both actual choices and elicited beliefs in Krupka and Weber (2013) are consistent with moral monotonicity.

7.3 Moral Monotonicity and Outside Options

Lazear et al. (2012) report an extended experimental design for dictator games that includes an outside option that allows subjects to opt out of the dictator game. The Lazear et al. Experiment 2 is a within-subjects design including several decisions with one selected randomly for payoff. In Decision 1, subjects play a distribute \$10 dictator game. In Decision 2, subjects can sort out of the \$10 dictator game, and be paid \$10 (with the other subject getting \$0), or sort in and play the distribute \$10 dictator game. In other decision tasks, subjects can sort out of a \$*S* dictator game, and be paid \$10 (with the other subject getting \$0), or sort in and play the distribute \$*S* dictator game. Values of *S* varied from 10.50 to 20.[16]

Explaining behavior of subjects in Experiment 2 who sorted into a $S > 10$ dictator game and kept more than 10 for themselves is straightforward. A more interesting behavior is that many subjects sorted out, and were paid 10, when they could have sorted into a $S > 10$ dictator game and retained more than 10 for themselves (and/or more than 0 for the other). For example, in the $S = 11$ game, the outside option pays (dictator, other) payoffs (10,0) whereas Pareto-dominating payoffs such as (11,0), (10.50, 0.50) and (10,1) are available to a subject who sorts into the dictator game. The reluctant/willing sharers model developed by Lazear et al. (2012) is consistent with behavior patterns in the experiment. That model is a utility function with three arguments: own payoff, other's payoff, and a binary indicator variable with value 1 for the sharing (dictator game) environment and value 0 for the non-sharing (outside option) environment. This type of behavior is consistent with our moral monotonicity model, in which choosing the outside option allows the decision maker to avoid moral costs from making the sharing decision whereas choosing to play the game involves this cost, as we now explain.

A subject has the right to choose the ordered pair of payoffs (10,0) by sorting out. When sorting out, the feasible choice set is $X^o = \{(10,0)\}$, and the moral reference point is $r^o = (10,0)$

[16] The experiment included anonymity and no-anonymity treatments.

(as the minimal expectations point and the initial endowment are both (10,0)). When choosing to sort in, the dictator makes a decision on distribution of S_j in treatment j, that is $X^{in} = \{x : x_1 + x_2 = S_j, x_i \geq 0, i = 1,2\}$. Since the dictator's sharing options include 0 and S_j, the minimal expectations point for the two-stage game is (0,0), which together with initial endowment (10,0) results in a moral reference point, r^{in} that is less favorable to the dictator. An example of choices consistent with moral reference point monotonicity, can be captured by maximization of $w_1(r)u(m) + w_2(r)u(y)$ for some increasing function $u(\cdot)$ and weights, $w_i(\cdot)$, as in the example in section 3.2.d. The model is consistent with behavior by an agent who chooses the (10,0) outside option rather than sorting in to play for some distribute S (> 10) dictator game with feasible payoffs that Pareto-dominate (10,0) contained in its opting in budget set (see Appendix E for details).[17]

7.4 Moral Monotonicity and Earned Endowments

Oxoby and Spraggon (2008) report an experiment with dictator games that includes treatments in which initial endowments are determined in a first stage. In the receiver earnings treatment, the recipient determined the initial endowment by their performance on a test that used 20 questions from the Graduate Management Admissions Test (GMAT) or the Graduate Record Examination (GRE). Depending upon the number of questions answered correctly, the recipient was provided an initial endowment of either CAN $10, CAN $20, or CAN $40. In the second stage, the dictator decided how much of this endowment they would like to take from the recipient. The dictator earnings treatment differed along two dimensions. First, the initial endowment was earned by the dictator's performance on the 20 question exam. Second, the dictator's decision in the second stage was to determine how much of the initial endowment they would like to give to the recipient.

Across both versions of the game, the minimal expectations point is (0,0). Therefore, as in the Korenok et al. (2014) experiment, changes in moral reference points across the two treatments are entirely determined by changes in endowment. Focusing on pairs for whom the initial endowment is CAN $40, the endowment component of the moral reference point is (0,40) in the receiver-earning treatment and (40,0) in the dictator-earning treatment. MMA would thus predict that the amount allocated to the recipient under the recipient earnings treatment is greater

[17] Appendix E provides numerical examples, such as sorting in when S=13 but sorting out when S=12.

than the amount allocated to the recipient under the dictator earnings treatment. Across all three wealth levels, the mean amounts allocated to recipients in the receiver-earning treatment are greater than the mean amounts allocated to recipients in the dictator-earning treatment, which is a pattern of results at odds with conventional rational choice theory but consistent with the predictions of moral monotonicity.

Korenok et al. (2017) extend this line of inquiry by adding a set of survey questions designed to elicit participants' feelings of ownership over the initial endowments. As in Oxoby and Spraggon (2008), treatments varied whether the initial endowment was earned by the recipient or dictator and the subsequent framing of the task as either give to or take from the recipient. Across all wealth levels, the mean amount allocated to the recipient under the recipient earnings treatment was greater than the amount allocated to the recipient under the dictator earnings treatment. Moreover, dictators felt a stronger sense of ownership over the endowment than did recipients in the dictator earnings treatment and vice versa in the receiver earnings treatment. Hence, both actual choices and feelings of ownership over endowments depend on property rights and initial allocations. Such data patterns are consistent with moral monotonicity, and highlight the importance of moral reference points.

8. IMPLICATIONS OF MORAL MONOTONICITY FOR PLAY IN GAMES WITH CONTRACTIONS

We next extend our discussion to illustrate the implications of moral monotonicity for choice in strategic games involving contractions. Games that have been studied in the previous literature include: (1) the moonlighting game and its contraction, the investment game, (2) carrot and stick games and a contraction in the positive domain (carrot game) as well as a contraction in the negative domain, (stick game). Together with dictator games, these games have been widely used in the literature to measure different aspects of social behaviors, including trust and cooperation. Moral monotonicity has different implications for play of these games than does conventional rational choice theory, or a stronger traditional assumption such as convex preferences (or GARP or social preference models).

8.1 Investment and Moonlighting Games

The investment game (Berg et al. 1995, and hundreds of other papers) can be constructed from the moonlighting game (Abbink et al. 2000, and scores of other papers) by contracting the feasible choice sets of the first and second movers to remove take options.[18] Conventional rational choice theory and moral monotonicity have different implications for such contractions and allow a way to distinguish between the two models using observed choice.

First, note that, for any given *positive* amount received, the second mover's (SM's) choice is the same in the moonlighting and investment games (with the same initial endowments). This is the prediction of conventional theory as well as moral monotonicity because the reference point for the SM opportunity sets is the same in the two games. Next, for any first mover (FM) who sends a non-negative amount in the moonlighting game, conventional theory requires that he choose the same amount to send in the investment game. Moral monotonicity, in contrast, requires him to choose a larger amount to send in the investment game, because the moral reference point for the FM opportunity set is more favorable to the FM in the moonlighting game than in the investment game; this is so because the initial endowments and the FM's minimal payoff are preserved but SM's minimal expected payoff is smaller in the moonlighting game than in the investment game (for details see Appendix F)

Existing data provide empirical support for moral monotonicity: We have analyzed data from an investment game experiment reported in Cox (2004) and a moonlighting game experiment reported in Cox, Sadiraj, and Sadiraj (2008). These two experiments used the same initial endowments $e = (10,10)$, the same multiplier k (=3) and were run by the same experimenter. Data from these experiments are consistent with the implications of moral monotonicity and inconsistent with the implications of conventional theory, as follows. We have data from 64 subjects who participated in the investment game and 130 subjects (66 within-subjects design and 64 between-subjects design) who participated in the moonlighting game.

FM choices: Using only FM data with non-negative amounts sent, we find that the means of the amounts sent are 5.97 (IG) and 4 (MG) and significantly different (t-test, p-value= 0.026.[19] Therefore, the FM data are consistent with the above implications of moral monotonicity but inconsistent with implications of conventional rational choice theory.

---

[18] In the moonlighting game (Abbink, et al. 2000), both players are endowed with the same amount of money. The first mover (FM) can give or take money from the second mover (SM); the maximum amount that can be given is the full endowment whereas the maximum amount that can be taken is one-half the endowment. Money given by FM is tripled by the experimenter but money taken is not transformed. After the SM is informed of the FM's choice, he/she can also give or take money from the FM. Each currency unit (CU) taken costs SM 1/3 CU whereas each CU given costs SM one CU. The investment game is a contraction in that FM and SM can only give and not take.

[19] If we examine only at Send > 0, averages are 7.35 (IG) and 4.84 (MG), which are significantly different (t-test, p-value=0.004) at conventional levels.

<u>*SM choices*</u>: Estimates (standard errors in parentheses) of censored regressions for SM choices at information sets with "FM not taking" (send ≥ 0, N=78) are[20]

$$E(r^s) = 0.67^{***}(\pm 0.15)\times s + 0.41(\pm 0.29)\times s\times D_M - 0.23(\pm 1.30)\times D_M$$

Insignificance of the coefficients for D_M and $s\times D_M$, "Moon" and "Send × Moon," are consistent with the (same) implication of moral monotonicity and conventional theory, as discussed above.

Taken jointly, we conclude that differences in play across the moonlighting and investment games are inconsistent with standard rational choice theory. Changes in the first mover's moral reference points across games leads to greater amounts shared in the investment game, a finding that is consistent with the predictions of moral monotonicity.

8.2 Carrot, Stick, and Carrot/Stick Games

Andreoni, Harbaugh and Vesterlund (2003) explore the effects of rewards and punishments on cooperation by studying behavior in three games: the carrot game that offers incentives only in terms of rewards, the stick game that allows only for negative incentives (punishment), and the carrot and stick (C&S) game that offers players both types of incentives. The two single incentive games are natural contractions of the C&S game.

Conventional theory and moral monotonicity predictions for SM's choice when the FM sends amount *s* are as follows (see Appendix F for details):

a. Carrot game: In this game the SM's choices can only increase the FM's monetary payoff by decreasing own monetary payoff. Conventional theory requires that if the SM choice in the C&S game is positive, then it remains a most preferred return in the carrot game. This is also the moral monotonicity prediction because the moral reference point does not change. As shown in their Figure 7, Andreoni et al. (2003) find larger demand for rewards in the C&S game than in the carrot game which is inconsistent with both conventional theory and moral monotonicity.

b. Stick game: In this game the SM's choices can only decrease the FM's monetary payoff by decreasing own monetary payoff. Conventional theory requires that if the SM's most preferred choice in the C&S game is to reduce the FM's monetary payoff then it remains a most preferred return in the stick game. Moral monotonicity, however, predicts in the stick

[20] Where send > 0 (N=64): $E(r^s) = 0.65^{***}(\pm 0.17)\times s + 0.42(\pm 0.36)\times s\times D_M - 0.14(\pm 1.87)\times D_M$

game a smaller return in absolute value because the moral reference point (through the minimal payoff) favors the SM. As shown in their Figure 6, Andreoni et al. (2003) report a result they characterize as “surprising” (pg. 898) that demand for punishment is larger in the C&S game than in the stick game. This result is inconsistent with conventional rational choice theory (and, in that sense, surprising). However, such differences are as predicted by moral monotonicity and thus not “surprising” when viewed through such lens.

In sum, received data from Andreoni et al. (2013) provides evidence inconsistent with standard rational choice theory and mixed support for moral monotonicity. Importantly, however, moral monotonicity can rationalize a data pattern that Andreoni et al. (2003) label as surprising − that the demand for punishment is greater in the C&S game than in the stick game. As the moral reference point for the SM in the stick game is more favorable than in the C&S game, this is what one would expect with moral monotonicity.

9. CONCLUDING REMARKS

When faced with the opportunity to share resources with a stranger, when and why do we give? The dictator game has emerged as a key data generator to provide researchers with a simple approach for eliciting other-regarding preferences in a controlled setting. The game has worked well in the sense that we now understand giving behaviors at a much deeper level. What has been less well explored is whether received results violate the basic foundations of economic theory.

Recent dictator game experiments reveal that choices of subjects in specific pairs of dictator games are inconsistent with conventional preference theory and social preferences models (List, 2007; Bardsley, 2008; Cappelen et al., 2013) and inconsistent with (more general) rational choice theory (Korenok et al., 2014) characterized by Sen’s (1971, 1986) Properties α and β. The designs of experiments that produce the anomalous data suggest how to extend rational choice theory to increase its empirical validity. The Korenok et al. (2014) experiment suggests that choices depend on endowments in ways not captured by conventional rational choice theory. Studies such as List (2007), Bardsley (2008), and Krupka and Weber (2013) suggest that choices depend on minimal payoffs in ways not captured by conventional rational choice theory.

In this spirit, we propose moral reference points and a restatement of Properties α and β that models dependence on them. An implication of this approach is preservation of the contraction properties of rational choice theory for feasible sets and subsets that have the same moral reference point. The variables that determine the moral reference points we propose are observable features of feasible sets – endowments and minimal expected payoffs (at other's maximum) payoffs. We report an experiment designed to test the central feature of the new theory: monotonicity of choice in the distinct (endowment and extreme payoff) dimensions of moral reference points. Data from the experiment largely reject prominent alternative theories in favor of moral monotonicity theory.

We show how moral monotonicity can rationalize data from other types of dictator games in the literature and explain how the theoretical model has implications for play of strategic games involving contractions of feasible sets that differ from implications of conventional theory.

Our findings highlight the importance of revisiting standard models to explore the role of moral reference points in a broader array of settings. We provide an explanation of how moral monotonicity is predictive of received findings in a range of games designed to elicit social and cooperative behaviors. In this way, our results have both positive and normative import. For empiricists and practitioners, the results herein provide an indication that moral reference points can play an important role in welfare calculations and program evaluation.

REFERENCES

Abbink, Klaus, Bernd Irlenbusch, and Elke Renner, "The Moonlighting Game: An Empirical Study on Reciprocity and Retribution." *Journal of Economic Behavior and Organization* 42, (2000): 265-77.

Afriat, Sidney N., "The Construction of Utility Functions from Expenditure Data," International Economic Review, 8, no.1 (1967), 67–77.

Andreoni, James, and John H. Miller. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica* 70, no. 2 (2002): 737-753.

Andreoni, James, William Harbaugh, and Lise Vesterlund. "The Carrot or the Stick: Rewards, Punishments, and Cooperation." *American Economic Review* 93, no. 3 (2003): 893-902.

Arrow, Kenneth J. "Rational Choice Functions and Orderings." *Economica* 26, no. 102 (1959): 121-127.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. "Social preferences and the response to incentives: Evidence from personnel data." *The Quarterly Journal of Economics* 120, no. 3 (2005): 917-962.

Bardsley, Nicholas. "Dictator Game Giving: Altruism or Artefact?" *Experimental Economics* 11, no. 2 (2008): 122-133.

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10, no. 1 (1995): 122-42.

Bolton, Gary E., and Axel Ockenfels. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review* 90, no. 1 (2000): 166-193.

Cabrales, Antonio, Raffaele Miniaci, Marco Piovesan, and Giovanni Ponti. "Social preferences and strategic uncertainty: an experiment on markets and contracts." *The American Economic Review* 100, no. 5 (2010): 2261-2278.

Cappelen, Alexander W., Ulrik H. Nielsen, Erik Ø. Sørensen, Bertil Tungodden, and Jean-Robert Tyran. "Give and Take in Dictator Games." *Economics Letters* 118, no. 2 (2013): 280-283.

Çelen, Boğaçhan, Andrew Schotter, and Mariana Blanco. "On blame and reciprocity: Theory and experiments." *Journal of Economic Theory* 169 (2017): 62-92.

Chernoff, Herman. "Rational Selection of Decision Functions." *Econometrica* 22, no. 4 (1954): 423-443.

Cox, James C. "How to Identify Trust and Reciprocity." *Games and Economic Behavior* 46, no. 2 (2004): 260-281.

Cox, James C., Daniel Friedman, and Vjollca Sadiraj. "Revealed Altruism." *Econometrica* 76, no. 1 (2008): 31-69.

Cox, James C., Klarita Sadiraj and Vjollca Sadiraj. "Implications of Trust, Fear, and Reciprocity for Modeling Economic Behavior." *Experimental Economics* 11 (2008): 1-24.

Cox, James. C., and Vjollca Sadiraj. "On Modeling Voluntary Contributions to Public Goods." *Public Finance Review* 35, no. 2 (2007): 311-332.

Cox, James C., and Vjollca Sadiraj. "A Theory of Dictators' Revealed Preferences." Paper presented at the Biennial Social Dilemmas Conference, Rice University (2010).

Cox, James C., and Vjollca Sadiraj. "Direct Tests of Individual Preferences for Efficiency and Equity." *Economic Inquiry* 50, no. 4 (2012): 920-931.

Debreu, Gerard. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. Yale University Press. (1959).

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127, no. 1 (2012): 1-56.

Engel, Christoph. "Dictator Games: A Meta Study." *Experimental Economics* 14, no. 4 (2011): 583-610.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl. "Does Fairness Prevent Market Clearing? An Experimental Investigation." *Quarterly Journal of Economics* 108, no. 2 (1993): 437-459.

Fehr, Ernst, and Klaus M. Schmidt. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114, no. 3 (1999): 817-868.

Ferraro, Paul J., and Michael K. Price. "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment." *Review of Economics and Statistics* 95, no. 1 (2013): 64-73.

Fisman, Ray, Shachar Kariv, and Daniel Markovits. "Individual Preferences for Giving." *American Economic Review* 97, no. 5 (2007): 1858-1876.

Flage, Alexandre. "Taking Games: A Meta-Analysis," *Journal of the Economic Science Association* (2024): 1-24.

Fudenberg, Drew, and David K. Levine. "Fairness, Risk Preferences and Independence: Impossibility Theorems." *Journal of Economic Behavior & Organization* 81, no. 2 (2012): 606-612.

Galperti, Simone, and Bruno Strulovici. "A Theory of Intergenerational Altruism." *Econometrica* 85, no 4 (2017): 1175-1218.

Grossman, Philip J. and Catherine C. Eckel, "Giving vs. Taking for a Cause." *Economics Letters* 132 (2015): 28-30.

Herz, Holger and Dmintry Taubinsky. "What Makes a Price Fair? An Experimental Study of Transaction Experience and Endogenous Fairness Views." *Journal of the European Economic Association* (2017), forthcoming.

Hicks, John R. *Value and Capital*. Oxford: Clarendon Press. (1946).

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. "Fairness and the Assumptions of Economics." *Journal of Business* 59, no. 4 (1986): 285-300.

Kessler, Judd B. and Leider, Stephen. "Norms and Contracting." *Management Science* 58, no. 1 (2012): 62–77.

Kimbrough, Erik O., and Alexander Vostroknutov. "Norms Make Preferences Social." *Journal of the European Economic Association*, (2015).

Korenok, Oleg, Edward L. Millner, and Laura Razzolini. "Impure Altruism in Dictators' Giving." *Journal of Public Economics* 97, (2013): 1-8.

Korenok, Oleg, Edward L. Millner, and Laura Razzolini. "Taking, Giving, and Impure Altruism in Dictator Games." *Experimental Economics* 17, no. 3 (2014): 488-500.

Korenok, Oleg, Edward L. Millner, and Laura Razzolini. "Feelings of Ownership in Dictator Games." *Journal of Economic Psychology* 61, (2017): 145-151.

Kőszegi, Botond, and Matthew Rabin. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics* 121, no. 4 (2006): 1133-1165.

Krupka, Erin L. and Roberto A. Weber. "Identifying Social Norms using Coordination Games: Why does Dictator Game Sharing vary?" *Journal of the European Economic Association* 11, no. 3 (2013): 495-524.

Landry, Craig E., Andreas Lange, John A. List, Michael K. Price, and Nicholas G. Rupp. "Is a Donor in Hand Better Than Two in the Bush? Evidence from a Natural Field Experiment." *American Economic Review* 100, no. 3 (2010): 958-983.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber. "Sorting in Experiments with Application to Social Preferences." *American Economic Journal: Applied Economics* 4, no. 1 (2012): 136-163.

Levitt, Steven D., and John A. List. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives* 21, no. 2 (2007): 153-174.

List, John A. "On the Interpretation of Giving in Dictator Games." *Journal of Political Economy* 115, no. 3 (2007): 482-493.

Oxoby, Robert J., and John Spraggon. “Mine and Yours: Property Rights in Dictator Games.” *Journal of Economic Behavior and Organization* 65, no. 3-4 (2008): 703-713.

Roth, Alvin E. “Independence of Irrelevant Alternatives, and Solutions to Nash’s Bargaining Problem.” *Journal of Economic Theory* 16, no. 2 (1977): 247- 251.

Samuelson, Paul A. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press, 1947.

Sen, Amartya K. "Choice Functions and Revealed Preference." *Review of Economic Studies* 38, no. 3 (1971): 307-317.

Sen, Amartya K. "Social Choice Theory." in *Handbook of Mathematical Economics*, vol. III, ed. by K. J. Arrow and M. Intriligator. Amsterdam: North-Holland (1986).

Sen, Amartya K. “ Internal Consistency of Choice.” *Econometrica* 61, no. 3 (1993), 495-521.

Tversky, Amos, and Daniel Kahneman. “Loss aversion in Riskless Choice: A Reference-Dependent Model.” *The Quarterly Journal of Economics* 106, no.4 (1991): 1039-1061

Varian, Hal R., “The Nonparametric Approach to Demand Analysis.” *Econometrica* 50, no. 4 (1982): 945–973.

Zhang, Le, and Andreas Ortmann. "The Effects of Take-Option in Dictator-Game Experiments: A Comment on Engel’s (2011) Meta-study." *Experimental Economics* 17, no. 3 (2014): 414-420.

APPENDICES

Appendix A. Moral Reference Points in the Presence of N Players

Endowments for n agents will typically be specified, hence are observable. Identification of observable minimal expectations payoffs for $n \geq 2$ players can proceed as follows. Let y denote the vector of payoffs of n players. Let the feasible set be a finite set F. Let y_j^o be the maximum feasible payoff for player $j \in \{1, \ldots, n\}$, that is

$$y_j^o(F) = \max\{y_j \mid \boldsymbol{y} \in F\}$$

The minimal expectations point, $\boldsymbol{y}_*^F$ is defined as follows. For each player j, define player i's minimal expectation payoff with respect to j as

$$y_{*ij}^F = \min\{y_i \mid (\boldsymbol{y}_{-j}, y_j^o) \in F\}$$

Let $S_i = \{y_{*ij}^F : j \neq i)$ be the set of i's minimal expectation points. Naturally, player i expects her payoff to be no smaller than the smallest element in S_i; thus $y_{*i}^F = \min S_i$, which is the i^{th} element of the vector $\boldsymbol{y}_*^F$.

Appendix B. Implications of Properties α_R, β_R and M_R

Let Γ denote the set of all finite subsets of payoff vectors and let $\{C(\cdot \mid r) : r \in R^n\}$ denote a family of choice functions: for all $S \in \Gamma$, $C(S \mid r) = c(S, r)$ where $c(S, r)$ is a non-empty subset of *S*. Feasible sets, *S* are finite and moral reference point, s^r is the λ convex combination of the minimal expectation point, s_* of *S* and some initial endowment, *e* which is not required to be from *S*. We say that a binary relation $\succeq_r$ is constructed from $C(\cdot \mid r)$ if for all payoff vectors *x*, *y*

$$x \succeq_r y \text{ if } x \in c(S, r) \text{ for some set } S \in \Gamma \text{ that contains } \{x, y\} \text{ and } r = s^r$$

We say that choice function $C(\cdot \mid r)$ is constructed from a binary relation, $\succeq_r$ if for all $S \in \Gamma$

$$c(S, r) = \{x \in S \mid x \succeq_r y, \text{for all } y \in S \text{ and } r = s^r\}$$

Observation 1 (Weak Orders and Choices). The following statements hold

a. If $C(\cdot \mid r)$ satisfies Properties α_R and β_R then $\succeq_r$ is a weak order.

b. If $\succeq_r$ is a weak order then its choice function, $C(\cdot \mid r)$ satisfies Properties α_R and β_R.

PROOF.

Part a. Take any given r and any $C(\cdot|r)$ that satisfies Properties α_R and β_R. Construct the menu of binary (r-) relations $\succeq_r$, as stated above. That is, $x\succeq_r y$ if x is chosen from some set S when the moral reference point is r and S contains both x and y. For any given r, this binary (r-) relation is complete and transitive. Indeed, if $x\succeq_r y$ and $y\succeq_r z$, then by construction of $\succeq_r$, there exists some X such that $\{x,y\}\subseteq X$ and $x\in c(X,r)$ as well as some Y such that $\{y,z\}\subseteq Y$ and $y\in c(Y,r)$. Consider the scenario in which the feasible set is $\{x,y\}$ and the initial endowment is $e=(r-\lambda r_*)/(1-\lambda)$, where $(r_*)_i=\min\{x_i,y_i\}$, for all i. contains all finite subset. Since Γ contains all finite subsets of payoff vectors, $(\{x,y\},r)$ is a feasible decision problem, and by property α_R, $x\in c(\{x,y\},r)$. Similarly, $y\in c(\{y,z\},r)$. To show that $x\succeq_r z$ it suffices to show that $x\in c(Z,r)$ when the feasible set is $Z=\{x,y,z\}$ and the initial endowment is $e=(r-\lambda z_*)/(1-\lambda)$, where z_* is the minimal expectation point of Z. Verify that $z^r=r$. Proceeding with an indirect proof: suppose that $x\notin c(Z,r)$. But then $y\notin c(Z,r)$ because if $y\in c(Z,r)$ then by property α_R, $y\in c(\{x,y\},r)$ and property β_R requires $x\in c(Z,r)$. Similarly, $y\notin c(Z,r)$ implies $z\notin c(Z,r)$ and therefore, $c(Z,r)$ is empty, is a contradiction. Completeness of the constructed $\succeq_r$: for any given x and y, take $S=\{x,y\}$, $e=(r-\lambda s_*)/(1-\lambda)$ and apply $c(S,r)$ being non-empty.

Part b: Let a menu of r-relation, $x\succeq_r y$, transitive and complete for any given r, be given. For any set, S and moral reference point, s^r construct

$$c(S,r)=\{s^*\in S\,|\,r=s^r \text{ and } s^*\succeq_r s \text{ for all } s\in S\} \qquad (*)$$

The choice set, $c(S,r)$ is not empty as $\succeq_r$ is complete and S is finite. To show properties α_R and β_R take any given sets F and G such that $F\subseteq G$ and $f^r=g^r$. To simplify notation, let r denote the common reference point, $f^r=g^r$.

Property α_R: If $x\in F\cap c(G,r)$ then, by construction (*), $x\succeq_r g$ for all $g\in G$ which together with $F\subseteq G$ implies that $x\succeq_r f$ for all $f\in F$, and therefore by (*) $x\in c(F,r)$.

Property β_R: Take any $x,y\in c(F,r)$ and suppose that $x\in c(G,r)$. By construction (*), $x\succeq_r g$ for all $g\in G$ and $y\succeq_r x$ as $x,y\in c(F,r)$. By transitivity of $\succeq_r$, $y\succeq_r g$ for all $g\in G$, hence $y\in c(G,r)$.

Observation 2 (Moral Monotonicity (MM)). We provide a proof for statement (b) in Moral Monotonicity (MM). Proof for statement (a) is similar. Suppose that $F \subseteq G$ and $f_i^r > g_i^r$ and $c(G, g^r) \cap F \neq \phi$. First, we prove the weak version of statement (b), that is choice monotonicity when only one dimension of the moral reference point changes, and then prove the general case of statement (b).

1. (Special Case) If $f_j^r = g_j^r$ then Property α_R and M_R imply that for all $g^* \in c(G, g^r) \cap F$ there exists $f^* \in c(F, f^r)$ such that $f_i^* \geq g_i^*$
2. If $f_j^r \leq g_j^r$ then Property α_R, β_R, M_R, and Pareto efficiency implies that for all $g^* \in c(G, g^r) \cap F$ there exists $f^* \in c(F, f^r)$ such that $f_i^* \geq g_i^*$

PROOF.

Part 1. Take any arbitrary, $g^* \in F \cap c(G, g^r)$. Consider the feasible set, F and the initial endowment $e = (g^r - \lambda f_*)/(1-\lambda)$ where g^r is the moral reference point in G and f_* is the minimal expectation point in F. With this initial endowment, the decision problem is choice from (F, g^r), and by Property α_R, $g^* \in c(F, g^r)$, which implies

$$g_i^* \leq \sup c_i(F, g^r) \leq \sup c_i(F, f^r)$$

where the second inequality follows Property M_R, (that is, $c_i(F, f^r) > c_i(F, g^r)$ as $f_i^r > g_i^r, f_j^r = g_j^r$. Since F is finite, there exists some $f^* \in c(F, f^r)$ such that $f_i^* = \sup c_i(F, f^r)$.

Part 2. Let $F \subseteq G$ and $f_i^r > g_i^r, f_j^r \leq g_j^r$. We first show that the statement is true when $F = G$, and then extend it to subsets F to conclude the proof. All proofs below are written for n=2 (i.e., two-players).

<u>Case 1</u>. $F = G$. By contradiction, suppose that there exists some $g^o \in c(F, g^r)$ such that $f_i^* < g_i^o$ for all $f^* \in c(F, f^r)$. As $c(F, f^r)$ is finite

$$\sup_{c(F,f^r)} f_i^* < g_i^o \qquad \text{(A.1)}$$

Consider choice from F when the initial endowment is $e = ((g_i^r, f_j^r) - \lambda f_*)/(1-\lambda)$. It is easy to verify that the moral reference point is $h = (g_i^r, f_j^r)$. Apply property M_R to:

(1) (F, f^r) and (F, h) to get $\sup_{c(F,h)} h_i^* \leq \sup_{c(F,f^r)} f_i^*$, and by (A.1)

$$h_i^* < g_i^o \text{ , for all } h^* \in c(F,h) \quad \text{(A.2)}$$

(2) (F,h) and (F,g^r) to get

$$\inf_{c(F,h)} h_j^* \le \inf_{c(F,g^r)} g_j^* \le g_j^o \quad \text{(A.3)}$$

Let $h^o \in c(F,h)$ be a point from $c(F,h)$ where *j*'s payoff is the smallest. By (A.2) and (A.3)

$$h_j^o \le g_j^o \text{ and } h_i^o < g_i^o$$

By property α_R, $h^o \in c(\{h^o,g^o\},h)$, by efficiency $g^o \in c(\{h^o,g^o\},h)$, by property β_R, $g^o \in c(F,h)$, but then (A.2) requires $g_i^o < g_i^o$ which is a contradiction.

Case 2. $F \subset G$. Take any arbitrary, $g^* \in c(G,g^r) \cap F$. Consider choice from F with initial endowment $e = (g^r - \lambda f_*)/(1-\lambda)$. The moral reference point is $g^r = (1-\lambda)e + \lambda f_*$, so by Property α_R, $g^* \in c(F,g^r)$. Apply Case 1 to choice from (F,g^r) and (F,f^r) to conclude that there exists some $f^* \in c(F,f^r)$ such that $g_i^* \le f_i^*$.

Appendix C. Example of a Choice Function that Satisfies Properties α_R, β_R and M_R

Without any loss of generality use "1" to denote dictator's index of the moral reference point. Dictator's valuation of final money allocation $x \in R_+^n$ when the moral reference point is $r \in R_+^n$ is

$$U(x,r) = \sum_{j=1..n} w_j(r)u(x_j) \quad \text{(A.4)}$$

for some positive increasing $u(\cdot)$ and weights, $w(\cdot)$ such as

$$\begin{aligned} w_j(r) &= \theta(kr_j)/M(r), \quad \text{if } j=1 \\ &= \theta(r_j)/M(r), \quad \text{otherwise} \end{aligned}$$

where $k \ge 1$ (captures "egocentricity" as in Cox and Sadiraj (2007, 2012)), $\theta(\cdot)$ is positive and increasing function and $M(r) = \theta(kr_1) + \sum_{j>1} \theta(r_j)$ (so the sum of weights is 1). Specification (A.4) says that individual's objective function is a weighted average of utility of own and other's final (money) payoffs with weights depending on the reference point.

The first two properties, α_R and β_R are clearly satisfied for any given reference point. To verify Property M_R, take any $F = G$ and $f_i^r > g_i^r$, $f_{-i}^r = g_{-i}^r$. Suppose that $f^* \in c(F,f^r)$ (i) and $g^* \in c(G,g^r)$ (ii). We show that $f_i^* \ge g_i^*$. It follows from (i) and (ii) that

$$(1)\quad w_i(f^r)u(f_i^*)+\sum_{j\neq i} w_j(f^r)u(f_j^*) \geq^{(i)} w_i(f^r)u(g_i^*)+\sum_{j\neq i} w_j(f^r)u(g_j^*)$$

$$(2)\quad w_i(g^r)u(g_i^*)+\sum_{j\neq i} w_j(g^r)u(g_j^*) \geq^{(ii)} w_i(g^r)u(f_i^*)+\sum_{j\neq i} w_j(g^r)u(f_j^*)$$

Case 1. $i=1$. Multiply both sides of (1) by $M(f^r)$ and (2) by $M(g^r)$ to get

$$(1a)\quad \theta(kf_1^r)u(f_1^*)+\sum_{j>1}\theta(f_j^r)u(f_j^*) \geq \theta(kf_1^r)u(g_1^*)+\sum_{j>1}\theta(f_j^r)u(g_j^*)$$

$$(2a)\quad \theta(kg_1^r)u(g_1^*)+\sum_{j>1}\theta(g_j^r)u(g_j^*) \geq \theta(kg_1^r)u(f_1^*)+\sum_{j>1}\theta(g_j^r)u(f_j^*)$$

Next, verify that

$$\theta(kf_1^r)u(f_1^*)+\sum_{j>1}\theta(f_j^r)u(f_j^*) \geq^{(1a)} \theta(kf_1^r)u(g_1^*)+\sum_{j>1}\theta(f_j^r)u(g_j^*)$$

$$=^{f_{-1}^r=g_{-1}^r} \theta(kf_1^r)u(g_1^*)+\sum_{j>1}\theta(g_j^r)u(g_j^*) \geq^{(2a)} \theta(kf_1^r)u(g_1^*)+\left[\theta(kg_1^r)u(f_1^*)+\sum_{j>1}\theta(g_j^r)u(f_j^*)-\theta(kg_1^r)u(g_1^*)\right]$$

$$=^{g_{-1}^r=f_{-1}^r} \theta(kf_1^r)u(g_1^*)+\left[\theta(kg_1^r)u(f_1^*)+\sum_{j>1}\theta(f_j^r)u(f_j^*)-\theta(kg_1^r)u(g_1^*)\right]$$

Drop the common term $\sum_{j>1}\theta(f_j^r)u(f_j^*)$ and rearrange terms to get

$$\theta(kf_1^r)u(f_1^*) \geq \theta(kf_1^r)u(g_1^*)+\left[\theta(kg_1^r)u(f_1^*)-\theta(kg_1^r)u(g_1^*)\right]$$

which is equivalent to

$$\left[\theta(kf_1^r)-\theta(kg_1^r)\right]\left(u(f_1^*)-u(g_1^*)\right) \geq 0$$

Inequalities $f_1^r > g_1^r$ and monotonicity of $\theta(\cdot)$ imply that the first term in the last statement is positive, and by monotonicity of $u(.)$ we have $f_1^* \geq g_1^*$.

Case 2: $i>1$. Follow the same steps as above to get

$$\left[\theta(f_i^r)-\theta(g_i^r)\right](u(f_i^*)-u(g_i^*)) \geq 0$$

which together with $f_i^r > g_i^r$, monotonicity of $\theta(\cdot)$ and $u(.)$ imply $f_i^* \geq g_i^*$.

C. 1. A numerical application to Give and Take Action sets in Dictator games in Inequality treatment.

Let the choice set be determined by maximization of $U(z\,|\,r)=\sum_{i=1,2} w_i(r)u(z_i)$, where

$w_1(r)=\dfrac{1.1^{2r_1}}{1.1^{r_2}+1.1^{2r_1}}$, $\mathrm{w}_2(r)=\dfrac{1.1^{r_2}}{1.1^{r_2}+1.1^{2r_1}}$ and $u(\cdot)$ some differentiable increasing concave function.

The feasible set is $S=\{z \mid z_1+z_2=30, z_1\in[19,27]\}$ and the moral reference point is $r=\lambda s_* +(1-\lambda)e$, where e is the ordered pair of initial endowments and $s_*=(19,3)$ is the minimal expectation payoff point.

Give-Inequality game. The initial endowment is (27,3) in Give-Inequality, so the moral reference point, is $r^g=(27-8\lambda,3)$.

Take-Inequality game. The initial endowment is (19,11) in Take-Inequality, so the moral reference point, $r^t=(19,11-8\lambda)$ is less favorable for the dictator for all $\lambda\in[0,1)$.

Compare choices in Give-Inequality and Take-Inequality treatments. If $\lambda=1$ then the moral reference point is simply the minimal expectation payoff point, (19,3). So by Property M_R, the choice set in Give-Inequality and Take-Inequality is the same. If $\lambda\in[0,1)$, then by Property M_R choice in Give-Inequality is more favorable to the dictator than in Take-Inequality, that is, $z_1^t<z_1^g$ because the moral reference point is more favorable to the dictator in Give-Inequality. It is straightforward to verify that this is true for the specification, U(.) above. Substituting $z_2=30-z_1$ in dictator's decision problem we get

$$\max_{z_1}\{w_1(r)u(z_1)+w_2(r)u(30-z_1) \mid z_1\in[19,27]\}$$

By concavity of $u(\cdot)$, f.o.c. is also sufficient so dictator's choice in this game, is determined by

$$F(z_1^*)=\mu u'(z_1^*)-u'(30-z_1^*)=0$$

where $\mu=w_1(r)/w_2(r)$. Note that $\mu^g>\mu^t$ as $\mu^g/\mu^t=1.1^{k(r_1^g-r_1^t)-(r_2^g-r_2^t)}=1.1^{8(k+1)(1-\lambda)}>1$, for all $\lambda\in[0,1)$. At the Give-Inequality optimal choice, z_1^g one has

$$F(z_1^g \mid Take)=\mu^t u'(z_1^g)-u'(30-z_1^g)<\mu^g u'(z_1^g)-u'(30-z_1^g)=0$$

That is, $z_1^g(.)$ is too large to be optimal in the Take-Inequality game.

Appendix D. Summary of Demographics

Each subject in the experiment participated in only one budget-line treatment.

Budget Line (Nr of Subjects)	Male	African-American	**Major**[a] (Social Sciences, STEM, other)	**Study Year** (Freshman, Junior, Sophomore, Senior)	**GPA**[b] Mean {st.dev}

Inequality (112)	34.8%	51.8%	(45, 29, 26)%	(29, 24, 29, 17)%	3.32 {0.52}
Equal (93)	35.5%	55.9%	(41, 25, 34)%	(29, 25, 25, 22)%	3.31{0.44}
Envy (101)	41.6%	47.5%	(42, 33, 25)%	(22, 24, 33, 20)%	3.36 {0.45}
All (306)	37.3%	51.6%	(43, 29, 28) %	(26, 24, 29, 19)%	3.33 {0.47}

Notes. [a]STEM includes actuarial science, biology, chemistry, computer science, engineering, math, neuroscience, physics; Social Sciences includes accounting, business, economics, education, finance, history, marketing, political science, psychology, sociology. [b]GPA is 2 (if below 2.24), 3 (2.25 to 3.24), 3.5 (3.25 to 3.74), 4 (3.75 to 4).

Appendix E. An Example of MM Choices in Dictator Games with Outside Options (Lazear et al. (2012) experiment)

Refer to A.4. If the subject sorts out, the feasible set is {(10,0)}, the endowment is e = (10,0), so the moral reference point is also r^o = (10,0). Hence,

$$V^{out} = w_1(r^o)u(10) + w_2(r^o)u(0)$$

If the player sorts in, then the feasible set contains all feasible allocations from dividing S. The minimal expectations point is (0,0) and given the initial endowment, (10,0) we get $r_1^{in} < 10 = r_1^{out}$ and $r_2^{in} = 0 = r_2^{out}$. Hence, $\theta(kr_1^{in}) < \theta(kr_1^{out})$ and $\theta(r_2^{in}) = \theta(r_2^{out})$, implying (*) $w_1(r^{out}) > w_1(r^{in})$. The decision-maker's problem is

$$V^{in}(z) = \max\{w_1(r^{in})u(z_1) + w_2(r^{in})u(z_2) : z_1 + z_2 = S\}$$

where z is the final payoff vector. Lazear et al. argue that because S > 10 an individual should sort in as (S,0) dominates (10,0), so (S,0) is preferred to (10,0). But this is not necessarily true for "moral cost" theory. Indeed, normalizing u(0)=0,

$$V^{in}(S,0) = w_1(r^{in})u(S) < w_1(r^{out})u(10) = V^{out}(10,0)$$

if

$$\frac{u(S)}{u(10)} \in \left(1, \frac{w_1(r^{out})}{w_1(r^{in})}\right)$$

Note that $w_1(r^{out}) / w_1(r^{in})$ does not depend on S whereas the ratio on the left hand side increases in S, therefore $V^{in}(S,0)$ will eventually exceed $V^{out}(10,0)$. In general, the subject will prefer sorting out if

$$V^{in}(z^*) = w_1(r^{in})u(z_1^*) + w_2(r^{in})u(S - z_1^*) < V^{out}(r^{out}) = w_1(r^{out})u(10).$$

Numerical illustrations. For a numerical illustration, let $u(z_i) = \sqrt{z_i / 10}$, $\lambda = 0.5$ and the weight specification in Appendix C.1,

$$w_1(r) = \frac{1.1^{kr_1}}{1.1^{r_2} + 1.1^{kr_1}},\ w_2(r) = \frac{1.1^{r_2}}{1.1^{r_2} + 1.1^{kr_1}}$$

An individual with $k = 2$, prefers staying out to sorting in when $S = 12$ but "sorts in" when $S = 13$ and keeps more than 10 as

$$V^{out}(10,0) \approx 0.87 > 0.85 \approx V^{in}(z^* = (10.4, 1.6)) \quad if\ S = 12$$
$$V^{out}(10,0) \approx 0.87 < 0.88 \approx V^{in}(z^* = (11.3, 1.7)) \quad if\ S = 13$$

An individual with k=1.1 "sorts in" when S =11 or S =12 but keeps less than 10 in both cases as

$$V^{out}(10,0) \approx 0.74 < 0.77 \approx V^{in}(z^* = (8.15, 2.85)) \quad if\ S = 11$$
$$V^{out}(10,0) \approx 0.74 < 0.80 \approx V^{in}(z^* = (8.89, 3.11)) \quad if\ S = 12$$

Appendix F. Play in Games with Contractions

F.1 Investment and Moonlighting Games

Let e denote the endowment of the FM and the SM. The amount sent by the FM is denoted by s. If s is positive, then it is multiplied by k >1 to obtain the amount received by the SM. Taking is not feasible in the investment game. In the moonlighting game, if s is negative then the multiplier is 1 to obtain the amount taken from the SM. The amount returned by the SM is denoted by r. Returning a negative amount is not feasible in the investment game. In the moonlighting game, when r is negative it costs the SM r/k to take r from the FM.

<u>SM opportunity sets across the two games</u>: Let the SM be in information set M_s for some *non-negative* amount s sent by the FM in the moonlighting game. The M_s set contains costly options for the SM but can increase/decrease FM's monetary payoff: $M_s = M_s^+ \cup M_s^-$ where

$$M_s^+ = \{(e - s + r, e + ks - r) : r \in [0, ks]\}$$
$$M_s^- = \{(e - s + r, e + ks + r / k) : r \in [-(e - s), 0)\}$$

Consider the SM's choice in M_s in the moonlighting game when the FM sends a non-negative amount. Consistent with observed behavior[21] (and Pareto efficiency), the amount returned will be from M_s^+.

What are alternative predictions for SM's choice in the investment game, at information set I_s given the same nonnegative s? In the investment game, the SM's choices can only increase the

[21] In data reported by Cox, Sadiraj, and Sadiraj (2008), only 2 out of 46 second movers who did not have money taken from them by first movers chose $r_s \in M_s^-$.

FM's monetary payoff by decreasing own monetary payoff,

$$I_s = \{(e-s+r, e+ks-r) : r \in [0, ks]\}$$

Thus $I_s = M_s^+ \subset M_s$. Conventional theory requires the same $r_s \in M_s^+$ to be the SM's choice in the investment game. This is also the moral monotonicity prediction because sets M_s and I_s have the same moral reference point.

__*FM choices across the two games*__: In the moonlighting game, the FM can send money to the SM or take up to one-half of the SM's initial endowment. Any positive amount sent ($s > 0$) is multiplied by $k > 1$. Any amount taken ($s < 0$) is not transformed (it is one for one). The FM choice set is $M = M^+ \cup M^-$ where

$$M^+ = \{(e-s, e+ks) : s \in [0, e]\}$$
$$M^- = \{(e-s, e+s) : s \in [-e/2, 0)\}$$

Suppose the FM's choice in the moonlighting game is some non-negative s_M. In the investment game, the FM can only send money to the SM. So, $I = M^+ \subset M$ as the FM choice set is

$$I = \{(e-s, e+ks) : s \in [0, e]\}$$

Conventional theory requires the non-negative amount s_M to be the FM's choice in the investment game also if it is chosen in the moonlighting game because the feasible set in the investment game is a contraction of the feasible set in the moonlighting game. In contrast, moral monotonicity says that the FM will send more in the investment game because the moral reference point in set *I* is more favorable to the SM than is the moral reference point in set M. This is so, because the initial endowments are the same in the two games but minimal expectation payoffs are: 0 for the FM and e for the SM in set I, but in the set M, they are 0 for the FM and e/2 for the SM. Hence, compared to set M, minimal expectation payoffs is more favorable to the SM in set I.

__*Implications for game play*__: Both conventional theory and moral monotonicity imply that, for any *positive* amount received, the SM's choices in the moonlighting and investment games are identical. We distinguish between two types of FMs: the ones who send in the moonlighting game and the ones who take. For a FM who takes in the moonlighting game, by design of the two games the FM must choose a larger amount in the investment game. For a FM who does not take in the moonlighting game, we have shown above that conventional theory predicts the same amount being sent in the two games whereas moral monotonicity predicts a larger amount being sent in the investment game.

F.2 Carrot, Stick, and Carrot/Stick Games

Let e = (240,0) in cents denote the endowments of the FM and the SM. The amount sent, s by the FM is the amount received by the SM and can take values from [40, 240] in all three games. The return, r_s by the SM can be positive (carrot), negative (stick) or either (C&S game) as returning a negative amount is not feasible in the carrot game whereas returning a positive amount is not feasible in the stick game. Despite the sign of the amount returned, the FM receives $5r_s$.

<u>*SM choices across the three games*</u>: For the amount s sent by the FM let the SM feasible sets be denoted by M^s_{cs} in the C&S game, M^s_c in the carrot game and M^s_s in the stick game such that $M^s_{cs} = M^s_c \cup M^s_s$. The M^s_{cs} set consists of options that are all costly for the SM but can increase or decrease FM's monetary payoff. The sets are:

$$M^s_c = \{(240 - s + 5r, s - r) : r \in [0, s]\}$$
$$M^s_s = \{(240 - s + 5r, s + r) : r \in [\max\{-(240 - s)/5, -s\}, 0]\}$$

Let r_{cs} be the SM's choice in the C&S game when the FM sends amount s. Conventional theory and moral monotonicity predictions for SM's choice when the FM sends amount s are as follows:

Carrot game: In this game the SM's choices can only increase the FM's monetary payoff by decreasing own monetary payoff. Conventional theory requires that if the SM choice in the C&S game is positive, i.e. $r_{cs} \in M^s_c$ then it remains a most preferred return in the carrot game. This is also the moral monotonicity prediction because sets M^s_{cs} and M^s_c have the same moral reference point as the initial endowments are the same and the two sets have the same minimal expectation point, 240-s for the FM and 0 for the SM.

Stick game: In this game the SM's choices can only decrease the FM's monetary payoff by decreasing own monetary payoff. Conventional theory requires that if the SM's most preferred choice in the C&S game is to reduce the FM's monetary payoff, i.e., $r_{cs} \in M^s_s$ then it remains a most preferred return in the stick game. Moral monotonicity, however, predicts in the stick game a smaller return in absolute value because the moral reference point favors the SM. Verify that the minimal expectation payoff for the FM is 240-s in C&S and Stick game, but for the SM is s in the Stick game and 0 in the C&S game.