Costly Switching from a Status Quo[*]

Begum Guney[†] Michael Richter[‡]

August 12, 2018

Abstract

We axiomatically characterize a theory of status quo-dependent choice where an agent faces switching costs that depend upon both the status quo and the alternative he switches to. In a choice problem with a status quo, the agent chooses the alternatives that yield the highest utility net of switching cost. This generates status quo bias and also allows for a wide range of reference effects. We examine the behavior of such agents in Prisoner's Dilemma (PD) games. In a single PD game, switching costs can lead to cooperation. However, across different PD games, it is not "anything goes" and instead we derive necessary and sufficient conditions for cooperation rates to be consistent with our model. We then verify that these conditions are satisfied by Charness et al.'s (2016) experimental data. We also perform a similar analysis for other theories such as models of status quo bias, magical thinking, inequity aversion, and fairness; and find that these theories make either invalidated or looser predictions.

Keywords: Choice; Switching Cost; Status Quo Bias; Reference Effect; Prisoner's Dilemma; Cooperation

JEL classification: D00; D01; D91; C72

[*] We thank Levent Kockesen, Ariel Rubinstein, Rani Spiegler, and Roee Teper for their helpful comments and suggestions. The first author acknowledges financial support from the Scientific and Technological Research Council of Turkey, Project #113K370.

[†] Department of Economics, Ozyegin University. E-mail: begum.guney@ozyegin.edu.tr

[‡] Department of Economics, Royal Holloway, University of London. E-mail: michael.richter@rhul.ac.uk

1 Introduction

It has become widely accepted that a status quo may influence choices. In this paper, we characterize a new axiomatic choice model which is based upon switching costs from a status quo. Switching costs are frequently observed in real-life decision problems. For example, an agent faces *transaction costs* when changing banks as closing/opening accounts can be tedious and payments must be set up anew. A user faces *learning costs* when switching operating systems as it takes time and energy to learn a new system, and *compatibility costs* if the new operating system is not fully compatible with his existing hardware and software. Other examples of switching costs include *contractual costs* incurred when a customer breaks a long-term contract by switching providers and *moral costs* incurred when an agent deviates from a social norm.[1] Empirically, there is a substantial literature documenting the impact of switching costs in the airline (Carlsson and Löfgren (2006)), credit card (Ausubel (1991)) and phone service (Shi et al. (2006)) markets, and in pricing (Chevalier and Scharfstein (1996)).

We characterize a choice procedure where an agent has a utility function and a switching cost function. When there is no status quo, he simply chooses rationally by maximizing his utility function. When there is a status quo, choosing any alternative other than the status quo involves a switching cost and the agent chooses the alternatives that maximize his utility net of the switching cost. These switching costs are non-negative and may depend on both the status quo and the alternative the agent switches to.

In order to characterize this procedure, we incorporate exogenous status quos into the classical choice framework, as in the canonical model of Masatlioglu and Ok (2005). Following a revealed preference approach, we impose three axioms: (i) choices are rational across problems with a fixed status quo position; (ii) choices weakly exhibit status quo bias (Samuelson and Zeckhauser (1988))[2]; and (iii) choices are continuous.

Our model generalizes the status quo model of Masatlioglu and Ok (2005, 2014). There, a status quo narrows the decision maker's focus to a smaller subset by ruling out some alternatives.[3] In our setting, such

[1] Norms that are personal or social can serve as status quos and deviating from them can induce moral costs on agents. See Akerlof and Kranton (2005), Gazzaniga (2005), and Levitt and List (2007).

[2] Status quo bias refers to the phenomenon that an alternative is viewed more positively if it is the status quo. See also Knetsch (1989), Madrian and Shea (2001), and Choi et al. (2004).

[3] To see how it differs from models with exogenous reference points that are not necessarily status quos, refer to Tversky and Kahneman (1991), Bossert and Sprumont (2003), Munro and Sugden (2003) and discussions in Masatlioglu and Ok (2005, 2014).

an agent would declare that he faces unbearable switching costs to all the ruled-out alternatives and zero costs otherwise. Allowing for more general switching costs, our model is able to incorporate a wider range of *reference effects*[4], which sets it apart from Masatlioglu and Ok (2005, 2014) and other status quo models that follow it, such as Ortoleva (2010), Ok et al. (2014), and Riella and Teper (2014).[5] The advantage of the model is shown using the Prisoner's Dilemma games since it is able to explain Charness et al.'s (2016) data which the fundamental status quo bias model cannot. Thus, the generalized model we suggest is not solely of theoretical interest, but has practical applications as well.

Apart from decision theory, switching costs are employed in repeated game settings in which a player incurs a cost whenever he changes his strategy from one stage to another. That framework is developed by Klemperer (1987,1995), Beggs and Klemperer (1992), and Lipman and Wang (2000, 2009) and is used to study topics ranging from inflation-targeting to pricing decisions and entry/deterrence decisions (see Farrell and Shapiro (1989), Von Weizsäcker (1984), Libich (2008), and Caruana and Einav (2008)).

In this paper, we apply our switching cost model to one-shot Prisoner's Dilemma games. While cooperation is possible in the presence of switching costs in a single Prisoner's Dilemma game, we find that it is not "anything goes" across such games. We derive necessary and sufficient conditions for the cooperation rates across different Prisoner's Dilemma games and verify our model's predictions using Charness et al.'s (2016) experimental data. We also do a similar analysis to test the predictions of the models of status quo bias (Masatlioglu and Ok (2005, 2014)), inequity aversion (Fehr and Schmidt (1999)) and fairness (Rabin (1993)), and find that they are violated by Charness et al.'s (2016) data. Furthermore, the necessary and sufficient conditions we derive for the magical thinking model (Daley and Sadowski (2017)) are satisfied by Charness et al.'s (2016) data, but they are looser than those of the switching cost model. Thus, the switching cost model has the highest Selten (1991) score; it makes the tightest validated predictions.

[4] A status quo may affect choices even when itself is not chosen and a *reference effect* is the change brought about in the relative ranking of options due to the presence of a status quo. An analysis of Masatlioglu and Uler's (2013) data reveals that a significant portion of choices (31.66%) violate status quo irrelevance, the main axiom that we relax (see Section 2 and Appendix II). For non-status quo reference effects such as the attraction effect, see Huber et al. (1982) and Simonson (1989); the phantom effect, see Farquhar and Pratkanis (1992, 1993), Pettibone and Weddell (2000), and Guney and Richter (2015); and for a comparison of those effects, see Highhouse (1996).

[5] Unlike the status quo-based choice models in the literature, we allow for unavailable status quos. However, it is not the first model in the literature to allow for the unavailability of reference points. Kalai and Smorodinsky (1975) and Rubinstein and Zhou (1999) do so in a bargaining setting while Guney et al. (2018) and an earlier version of Masatlioglu and Ok (2014) do so in a choice theory setting.

The rest of the paper is organized as follows. Section 2 presents the decision theoretic analysis. In Section 3, Prisoner's Dilemma games are analyzed both theoretically and using the data of Charness et al. (2016). Concluding remarks appear in Section 4. All proofs are presented in Appendix I. The details regarding Masatlioglu and Uler's (2013) design and our analysis of reference effects therein are provided in Appendix II.

2 The Model

Define X to be a compact metric space that represents the grand set of alternatives. A *preference relation* $\succeq$ is a transitive and reflexive binary relation on X. A *choice problem* is a pair (S, x) where S is the set of *available* alternatives and x is the *status quo position*. The set S is a non-empty closed subset of X. The element x is either in X, in which case it is the status quo, or it is an object $\diamond$ that does not belong to X, interpreted as a no status quo situation. If the status quo $x \in S$, then it is *feasible* and if $x \in X \backslash S$, then it is *infeasible*. The set of all choice problems with a status quo is $\mathcal{C}_{sq}(X)$ while the set of those without a status quo is $\mathcal{C}_{\diamond}(X)$, and $\mathcal{C}(X)$ is their union. A choice correspondence is a non-empty valued map $c : \mathcal{C}(X) \rightrightarrows X$ such that $c(S, x) \subseteq S$ for any $(S, x) \in \mathcal{C}(X)$.[6]

To illustrate, consider a decision maker who receives a set O of job offers and must choose among them. If he is just out of college and unemployed, then he faces a choice problem without a status quo, i.e. $(O, \diamond)$. If he currently has a job, then his choice problem has a feasible status quo, i.e. $(O \cup \{x\}, x)$, since he can always choose to stick with his present job. Finally, he might have just lost his job and even though it is no longer available to him, the lost job may still affect his choice. In this case, he faces a choice problem with an infeasible status quo, i.e. (O, x).

In the classical theory, the Weak Axiom of Revealed Preferences (WARP) makes it possible to view choices as resulting from the maximization of a complete preference relation. Our first axiom extends WARP to choice problems with the same status quo position (either a fixed status quo, i.e. $x \in X$, or no status quo, i.e. $x = \diamond$).

[6] This notation follows that of Masatlioglu and Ok (2005). It is also consistent with Salant and Rubinstein (2008) who would refer to the status quo x as a "frame". Note that, in order to simplify the notation, we drop the curly brackets that would be used to denote sets within the choice correspondence. For instance, we write $c(xy, z)$ instead of $c(\{x, y\}, z)$.

Axiom 2.1 *(WARP) For any* $(S,x),(T,x) \in \mathcal{C}(X)$ *such that* $T \subseteq S$,

$$c(T,x) = c(S,x) \cap T, \text{ provided that } c(S,x) \cap T \neq \emptyset.$$

This axiom states that an agent is rational across problems with the same status quo and across problems without a status quo. Therefore, irrational behavior can only result from a change in the status quo and the introduction/removal of a status quo.

The following axiom specifies the choices that are vulnerable to the status quo bias phenomenon.

Axiom 2.2 *(Status Quo Bias (SQB)) For any* $x, y \in X$,

(i) if $y \in c(xy,\diamond)$, *then* $y \in c(xy,y)$

(ii) if $y = c(xy,\diamond)$, *then* $y = c(xy,y)$.

This axiom ensures that designating an alternative as the status quo doesn't harm its relative value. Equivalent versions of this axiom can be found in Masatlioglu and Ok (2014) and similar ones in Sugden (1999), Munro and Sugden (2003), Masatlioglu and Ok (2005), Sagi (2006), Apesteguia and Ballester (2009), Ortoleva (2010), and Riella and Teper (2014). It is this axiom that motivates the interpretation of a reference point in our model as a status quo.

Finally, the following technical axiom ensures that the agent makes similar choices across similar choice problems. It is vacuously satisfied when X is finite.

Axiom 2.3 *(Upper Hemi-Continuity (UHC)) Suppose that* $x, y, z \in X$ *and* (x_n), $(y_n) \in X^\infty$ *are convergent sequences such that* $x_n \to x$ *and* $y_n \to y$. *Then,*

(i) $y_n \in c(x_n y_n, x_n)$ $\forall n$ *implies* $y \in c(xy,x)$.

(ii) $y_n \in c(x_n y_n, \diamond)$ $\forall n$ *implies* $y \in c(xy,\diamond)$.

(iii) $y_n \in c(x y_n, z)$ $\forall n$ *implies* $y \in c(xy,z)$

Part (ii) states that an agent behaves continuously when there is no reference point, while part (iii) states that he behaves continuously given a fixed reference point.[7] Part (i) introduces a further requirement that the

[7] Rather than focusing on tuples, these parts could be phrased more generally and equivalently (given the other axioms) as $c(\cdot,\diamond)$ being upper hemi-continuous and $y_n \in c(S \cup y_n, z)$ implying $y \in c(S \cup y, z)$ for any $(S,z) \in \mathcal{C}_{sq}(X)$. It is straightforward to show that these formulations are equivalent to Masatlioglu and Ok's (2014) continuity axiom on their domain of available status quos.

effect of continuously changing reference points on choice is continuous, and this is a novel and intuitively appealing condition.

Below, our main theorem shows the equivalence between the three aforementioned axioms and a choice procedure of costly switching from a status quo.

Theorem 2.1 *A choice correspondence c satisfies Axioms WARP, SQB, and UHC if and only if there exist*

(1) a continuous function $U : X \to \mathbb{R}$,

(2) a lower semi-continuous function $D : X \times X \to \mathbb{R}_+$ with $D(x,x) = 0$ for any $x \in X$, such that

$$c(S, \diamond) = \operatorname*{argmax}_{s \in S} U(s) \quad \textit{for any } (S, \diamond) \in \mathcal{C}_{\diamond}(X)$$

$$c(S, x) = \operatorname*{argmax}_{s \in S} U(s) - D(s,x) \quad \textit{for any } (S, x) \in \mathcal{C}_{sq}(X).$$

The "if" direction is straightforward. Regarding the "only if" direction, a straightforward idea would be to consider status quo-dependent choices and to define rationalizing status quo-dependent preference relations $\succeq_x$ for any $x \in X$. Then, one would represent these preference relations as $\overline{U}(\cdot, x)$ and define $U(s) := \overline{U}(s,s)$ and $D(s,x) := -\overline{U}(s,x) + \overline{U}(s,s)$ for any $s, x \in X$. This argumentation is only halfway successful because although U and D together represent the agent's choices, the non-negativity of D is not guaranteed.[8] The bulk of the proof (presented in Appendix I) involves the creation of an *auxiliary preference relation* from which a function $\overline{U}$ is shown to exist with the property $\overline{U}(s,s) \geq \overline{U}(s,x),\ \forall s, x \in X$, thereby delivering the non-negativity of D.

The representation is based on two endogenously derived functions, U and D. The function U conveys the status quo-free utility derived from each alternative, while D reveals, for any $(y,x) \in X \times X$, the switching cost incurred when the agent chooses alternative y in the presence of the status quo x. As mentioned earlier, there are different types of switching costs, such as physical costs, transaction costs, and moral costs. Our model is applicable in all these cases since it does not tie the agent to a specific type of switching cost. We use the letter D to capture the idea of switching costs as a "decrement" in utility.

[8] This non-negativity is essential for both our interpretation and the "if" direction.

In the absence of a status quo, the decision maker behaves rationally by choosing the available alternatives that maximize U. However, when there is a status quo, say x, the agent chooses the available alternatives with the highest utility net of the switching cost, $U(\cdot) - D(\cdot, x)$.

For any status quo x, the cost function $D(\cdot, x)$ is minimized at x. This feature guarantees that whenever an alternative is (weakly) chosen over a status quo, its status quo-free utility is at least as high as the status quo option's utility. As a result, the choices characterized here are immune to money pump arguments (Echenique et al. (2011)). More precisely, for any distinct $x, y, z \in X$, the following choice cycle cannot be accommodated in our model.

$$c(xy, x) = y,\ c(yz, y) = z \text{ and } c(xz, z) = x.$$

Identification: As in most deterministic choice models, observed choices are sufficient for ordinal but not cardinal identification, and this is true here as well. More specifically, U and $U - D$ can be ordinally identified given the choice data $c(S, \diamond)$ and $c(S, z)$, respectively. However, D cannot be ordinally identified when it doesn't affect choices. For example, take $X = \{x, y, z\}$ such that $U(x) > 10 > 1 > U(y)$ and $U(x) - D(x, z) > U(y) - D(y, z)$. Then, $D(x, z)$ and $D(y, z)$ can take any value between 0 and 9 and so cannot be ordinally ranked. In this case, it cannot be determined whether it is costlier to switch from z to x or z to y and furthermore this has no effect on choice since these switching costs are not large enough to alter the agent's ranking of these alternatives.

We now turn to some special cases of the model. First, if $D(y, x) = 0$ for any $x, y \in X$, then the agent behaves rationally. Another important special case is one where the cost of choosing an alternative other than the status quo is either zero or a prohibitively large fixed amount. Such costs can be mathematically represented by an indicator function and this restricted model is identical to the Q-model of Masatlioglu and Ok (2014). It relies on an additional axiom called status quo irrelevance (given below for our setting).

Status Quo Irrelevance (SQI). *For any $x, y, z \in X$,*

(i) If $y \in c(xy, x)$ and $z \in c(xz, x)$, then $c(yz, \diamond) = c(yz, x)$.

(ii) If $c(xy, x) = x = c(xz, x)$, then $c(yz, \diamond) = c(yz, x)$.

According to this specialized model, if the cost of switching from x to y and from x to z are both zero (or both prohibitively large), then the presence of the status quo x does not affect the choice from $\{y, z\}$. Thus, the status quo x can generate a reference effect only when it prohibits exactly one of y and z.

In contrast, our model allows for "graded reference effects" whereby the status quo x may reverse the relative ranking of y and z *even* when the agent is willing to leave the status quo for each (i.e. neither y nor z is prohibited). To see this, consider an agent who faces three alternatives $\{0, 1, 2\}$ with $U(x) = 2.5x$ and $D(x, y) = |x - y|^2$. When there is no status quo, the agent ranks $2 \succ_\diamond 1 \succ_\diamond 0$. However, when the status quo is 0, the agent ranks $1 \succ_0 2 \succ_0 0$. The status quo 0 changes the relative ranking of 1 and 2 without forbidding either of them. This behavior cannot be accommodated in the Q-model.

Masatlioglu and Ok (2005) consider a further specialization where for each status quo x, an agent faces a threshold $\phi(x) > 0$ and will only consider other alternatives y for which $U(y) - U(x) \geq \phi(x)$. In our framework, this can be represented by $D(y, x) = \phi(x)$ whenever $x \neq y$. Thus, the relationship between the three models can be stated as follows:

$$\phi(x)\text{-model} \subsetneqq Q\text{-model} \subsetneqq \text{Switching Cost model}$$

While some real-life scenarios naturally fit either the $\phi(x)$-model or the Q-model, many important situations require the fully general model. For example, a move from Windows to a Mac or to a Linux system involves compatibility costs, since only some Windows programs are available for the Mac and even fewer for Linux. Moreover, the Mac and Linux systems differ in cost. These switching costs depend on both the status quo and the alternative the agent switches to. Therefore, they cannot be represented by a ϕ function that depends only on the status quo.

Remark 1: Notice that WARP guarantees the existence of preference relations $\{\succeq_x\}_{x \in X}$ and $\succeq$ that rationalize the status quo-dependent and status quo-free choices, respectively. Axiom SQB then ensures a link between these preferences (specifically, $x \succsim y \Rightarrow x \succsim_x y$ and $x \succ y \Rightarrow x \succ_x y$). This representation is equivalent to our switching cost model in terms of the choices it gives rise to, though it does not provide a direct route to our representation, as discussed immediately after Theorem 2.1. Furthermore, the model with switching costs has several advantages. First, it is intuitively appealing to explain the status quo bias through switching costs from a status quo. Second, the switching cost model is tractable for applications as exemplified by the Prisoner's Dilemma game (Section 3).

Remark 2: In our model, $D \geq 0$. If $D(y,x)$ were allowed to be negative when $y \succ x$, then all the axioms would continue to hold. In other words, the model in which an agent can receive a utility "increment" rather than a "decrement" when switching from a status quo to a higher-utility alternative would be equivalent to our model. On the other hand, if D were allowed to be negative when an agent switches to a lower-utility alternative, then this could lead to violations of Axiom SQB. While such a model would not be ideal for the characterization of status quo-based choices, it might still be of interest as a choice model based on a general reference point. That procedure with a possibly negative D is characterized by the WARP and UHC axioms. In this case, the halfway successful approach suggested immediately after Theorem 2.1 would now be fully successful because the non-negativity requirement on D is relaxed.

Remark 3: The two most common loss-aversion models in the literature are Tversky and Kahneman's (1991) general loss aversion and additive linear loss aversion models. These models satisfy WARP and UHC, and therefore nest into the switching cost model with a possibly negative D. Their additive linear loss aversion model additionally satisfies Axiom SQB and therefore nests within Theorem 2.1.[9] In these models, alternatives are multi-dimensional and each alternative may have multiple gains and losses relative to a reference point, which the agent must then aggregate to an overall cost function. In contrast, our switching cost model operates over an abstract space and the function D already gives the overall cost.

3 An Application to Prisoner's Dilemma Games

In this section, we analyze Prisoner's Dilemma games using various equilibrium concepts, starting with one based on switching costs. Formally, a standard one-shot *Prisoner's Dilemma* game is a 2×2 game where both players face the same action sets $A_1 = A_2 = \{Coop, Defect\}$ and have symmetric utility functions $U_1(x,y) = U_2(y,x)\ \forall x,y \in \{Coop, Defect\}$ such that $U_1(Defect, Coop) > U_1(Coop, Coop) > U_1(Defect, Defect) > U_1(Coop, Defect)$. Notice that $Defect$ is a dominant strategy and $(Defect, Defect)$ is the unique standard Nash equilibrium of this game.

[9] Masatlioglu and Ok (2014) show that the general loss aversion model need not satisfy Axiom SQB and therefore does not nest in Theorem 2.1.

3.1 Switching Costs

In a one-shot game, both $(Defect, Defect)$ and $(Coop, Coop)$ are possible equilibrium outcomes given appropriate status quos and sufficiently high switching costs. However, and perhaps surprisingly, it is not "anything goes" across a constant population and different Prisoner's Dilemma games. We show this by first defining a population equilibrium concept and then deriving conditions on cooperation rates across different Prisoner's Dilemma games. Finally, we test these conditions using Charness et al.'s (2016) data.

In a Prisoner's Dilemma game, an agent with a status quo and a switching cost function can be represented by an element of a type space $\Theta = \{Coop, Defect\} \times \mathbb{R}^2_+$. This motivates the following definition of a Prisoner's Dilemma D-game for a population who is randomly matched into pairs to play it.[10,11]

Definition: A *Prisoner's Dilemma D-Game* is a pair $< \Gamma, F >$ where Γ is a standard Prisoner's Dilemma game and F is a continuous distribution on Θ.

For a player's decision whether to cooperate or defect, all that matters is his belief about whether others are choosing $Coop$ or $Defect$. In equilibrium, all players have correct beliefs and this leads to the following definition of a population equilibrium.[12]

Definition: In a Prisoner's Dilemma D-game $< \Gamma, F >$, a *population equilibrium* is a measurable map $a : \Theta \to \{Coop, Defect\}$ such that for every $\theta = (s_\theta, D_\theta) \in \Theta$,

$$a(\theta) \in \underset{a \in \{Coop, Defect\}}{\operatorname{argmax}} \; p\, U_1(a, Coop) + (1-p)\, U_1(a, Defect) - D_\theta(a, s_\theta)$$

where $p := F(\{\theta \in \Theta : a(\theta) = Coop\})$.

Charness et al. (2016) ran a Prisoner's Dilemma experiment, denoted by Γ_x, where x is the same for both players and equals 3, 4, 5, or 6. Payoffs are shown in the following table.

[10] In the model, agents are randomly matched into pairs to play a Prisoner's Dilemma D-game. Given expected utility, it would be equivalent if each player plays with every other player and receives their average payoff.

[11] We express our thanks to an anonymous referee who suggested the simplifying notation and definition.

[12] Note that the equilibrium condition in the definition is phrased only in terms of the first player's utilities because agents face the same action sets and have symmetric utility functions.

	Coop	Defect
Coop	x, x	$1, 7$
Defect	$7, 1$	$2, 2$

Table 1: Charness et al.'s (2016) Prisoner's Dilemma games

For a fixed population F, we study the population equilibria of the Prisoner's Dilemma D-games $< \Gamma_x, F >$ and denote them as a_x. The following proposition shows that the population equilibrium a_x is well-defined and unique (up to a measure 0 set of ties) for each x. For a population equilibrium a_x, the fraction of agents who cooperate is denoted by p_x and referred to as the *cooperation rate*.

Proposition 3.1 *For each Prisoner's Dilemma D-game $< \Gamma_x, F >$ where F has full support and $x \leq 6$, there is a unique population equilibrium.*

The following proposition provides three necessary and sufficient conditions on p_x for a sequence of population equilibria to be derived from a fixed population F. First, it must be that p_x is increasing in x. Second, there is an upper bound on how quickly p_x can increase. Third, p_x is non-trivial, there are both cooperators and defectors.

Proposition 3.2 *Let $(p_x)_{x \in \{3,4,5,6\}}$ be a sequence of probabilities. The sequence (p_x) is a cooperation rate sequence for some Prisoner's Dilemma D-games $< \Gamma_x, F >$ where F has full support if and only if*

- *p_x is strictly increasing in x,* *(monotonicity condition)*
- $p_x < \left(\frac{7-x}{6-x}\right) \cdot p_{x-1}$ *for $x \in \{4, 5\}$,* *(threshold condition)*
- $0 < p_x < 1$, *for all x.* *(non-triviality condition)*

The non-triviality condition is straightforward, given a distribution with full support – some agents will cooperate and some will defect. The monotonicity condition is intuitive and was a hypothesis in Charness et al. (2016). On the other hand, the threshold condition is novel. To understand it, notice that an agent cooperates if and only if his status quo is $Coop$ and

$$(6 - x) \cdot p_x + 1 < D_\theta(Defect, Coop).$$

The left-hand side of the above inequality provides a cooperation threshold.[13] For more agents to cooperate when x increases, it must be that the cooperation threshold decreases. The threshold condition is equivalent to requiring that $(6-x)p_x+1$ is decreasing in x.

In Table 2 below, the observed cooperation rates are those from Charness et al. (2016). We further calculate the cooperation thresholds and derive bounds on the cooperation rates. To understand these derived bounds, consider $x=4$. By monotonicity, the cooperation rate p_4 must be $23.46\% \leq p_4 \leq 50.60\%$. By the threshold condition, it must be that $p_5 < (2/1)\cdot p_4$ and that $p_4 < (3/2)\cdot p_3$. Therefore, $25.30\% < p_4 < 35.19\%$. Consequently, $\max\{23.46\%, 25.30\%\} < p_4 < \min\{50.60\%, 35.19\%\}$ where in both cases the tighter bounds are due to the threshold condition. Bounds for the other p_x's can be derived similarly. Since both conditions imply bounds, the tighter bound is indicated with an m (monotonicity) or t (treshold) below.

x	Observed Coop Rate	$(6-x)p_x+1$	Coop Lower Bound	Coop Upper Bound
3	23.46%	1.70	$22.45\%^t$	$33.67\%^m$
4	33.67%	1.67	$25.30\%^t$	$35.19\%^t$
5	50.60%	1.51	$33.67\%^m$	$60.47\%^m$
6	60.47%	1.00	$50.60\%^m$	100%

Table 2: Cooperation rates from Charness et al. (2016) and bounds derived from Proposition 3.2

Table 2 demonstrates that the monotonicity and threshold conditions both hold, as must the derived bounds. In a structural estimation, these cooperation thresholds can be used to back out agents' switching costs and status quos. For example, when $x=3$, it is estimated that 23.46% of the population has a status quo of $Coop$ and switching costs higher than 1.70.[14]

3.2 Other Models that Lead to Cooperation

In this section, we examine other theories in the literature that can explain cooperation in Prisoner's Dilemma games and study their predictions with the goal of comparing them to our theory's predictions. Overall, we find that our theory provides the tightest predictions that are in line with Charness et al.'s (2016) data.

[13] For further details on this threshold, see Lemma 3.0 in Appendix I.

[14] Alternatively, 76.54% of the population has a status quo of $Defect$ or switching costs less than 1.70.

3.2.1 Q-Model (Masatlioglu and Ok (2005, 2014))

In the Q-model, an agent's type is expressed by his status quo ($Coop$ or $Defect$) and his constraint sets ($Q(Coop)$ and $Q(Defect)$). Any heterogeneous population of agents can be partitioned into three groups. The first group consists of all agents for whom $Defect$ is the status quo. The alternative $Defect$ is their highest-utility alternative and is available in their Q-set (because $x \in Q(x)$ for all x). Hence, these agents always defect. The second group consists of agents for whom $Coop$ is the status quo and $Defect \in Q(Coop)$. These agents also defect because $Defect$ is their best option and it is available in their Q-set. Finally, the third group consists of agents for whom $Coop$ is the status quo and $Q(Coop) = \{Coop\}$. These agents necessarily cooperate because $Coop$ is the only alternative that they consider. Therefore, for a given x, with payoffs from Table 1, the cooperation rate p_x expresses the fraction of agents in this third group. Given a fixed population, the size of the third group does not vary and therefore the cooperation rate is predicted to remain constant in x. In other words, Q-model predicts $p_3 = p_4 = p_5 = p_6$. However, this is not the case in Charness et al.'s (2016) data.

3.2.2 Magical Thinking (Daley and Sadowski (2017))

In the *magical thinking* model, a decision maker thinks that his opponent takes the same action as he does with probability λ and takes the equilibrium action with probability $1 - \lambda$. Hence, the type space is $\Theta^{MT} = [0, 1]$. We now define Prisoner's Dilemma games and population equilibrium for the magical thinking model, as we did for the switching cost model.

Definition: A *Prisoner's Dilemma MT-Game* is a pair $< \Gamma, F >$ where Γ is a standard Prisoner's Dilemma game and F is a continuous distribution on Θ^{MT}.

Definition: In a Prisoner's Dilemma MT-game $< \Gamma, F >$, an *MT-population equilibrium* is a measurable map $a : \Theta^{MT} \to \{Coop, Defect\}$ such that for every $\lambda \in \Theta^{MT}$,

$$a(\lambda) \in \underset{a \in \{Coop, Defect\}}{\text{argmax}} \quad \lambda \cdot U_1(a, a) + (1 - \lambda) \cdot (p \cdot U_1(a, Coop) + (1 - p) \cdot U_1(a, Defect))$$

where $p := F(\{\lambda \in \Theta^{MT} : a(\lambda) = Coop\})$.

For a fixed population F, we study MT-population equilibria of the Prisoner's Dilemma MT-games $< \Gamma_x, F >$. As before, there exists a unique MT-population equilibrium for each x. The fraction of agents who cooperate in an MT-population equilibrium is denoted as p_x and referred to as the *magical thinking cooperation rate*.

Proposition 3.3 *For each Prisoner's Dilemma MT-game $< \Gamma_x, F >$ where F has full support and $x \leq 6$, there is a unique MT-population equilibrium.*

In the spirit of revealed preference, Daley and Sadowski (2017) provide a characterization of their model over the domain of all simultaneous and symmetric 2×2 games. However, as is well-known in the revealed preference literature, the set of axioms which characterize a procedure over a large data set is generally insufficient for smaller domains. The following proposition offers a novel characterization of magical thinking cooperation rates for Charness et al.'s (2016) domain.

Proposition 3.4 *Let $(p_x)_{x\in\{3,4,5,6\}}$ be a sequence of probabilties.*
(i) *The sequence (p_x) is a magical thinking cooperation rate sequence for some Prisoner's Dilemma MT-games $< \Gamma_x, F >$ where F has full support if and only if*

- p_x *is strictly increasing in* x, *(monotonicity condition)*
- $p_x < \dfrac{(x-2)(7-x)}{(x-3)(6-x)} p_{x-1} + \dfrac{1}{(x-3)(6-x)}$ *for* $x \in \{4, 5\}$, *(threshold condition)*
- $0 < p_x < 1$, *for all* x. *(non-triviality condition)*

(ii) *The switching cost cooperation rate conditions are analytically tighter than those of magical thinking.*

As in the switching cost model, the conditions on cooperation rates here come from three sources: a non-triviality and a monotonicity condition both of which are the same as in our model, and a threshold condition that is different than ours. Part (ii) of the above proposition demonstrates analytically that any cooperation rate p_x that can be explained by our switching cost model can also be explained by the magical thinking model. However, the reverse does not necessarily hold since the magical thinking threshold condition is strictly weaker than the switching cost threshold condition.

Table 3 shows that for Charness et al.'s (2016) data, the cooperation rate bounds of the magical thinking model are strictly weaker than those of the switching cost model. In fact, for this data set, all of the magical thinking bounds come from the monotonicity condition. Thus, the switching cost model is strictly more restrictive because its threshold conditions are tighter. Consequently, the switching cost model has a higher Selten (1991) score than the magical thinking model.[15]

x	Observed Cooperation	Switching Cost Model		Magical Thinking Model	
	Rate	Lower Bound	Upper Bound	Lower Bound	Upper Bound
3	23.46%	22.33%^{t}	33.67%^{m}	0%	33.67%^{m}
4	33.67%	25.30%^{t}	35.19%^{t}	23.46%^{m}	50.60%^{m}
5	50.60%	33.67%^{m}	60.47%^{m}	33.67%^{m}	60.47%^{m}
6	60.47%	50.60%^{m}	100%	50.60%^{m}	100%

Table 3: Cooperation rates from Charness et al. (2016) and bounds derived from Propositions 3.2 and 3.4

3.2.3 Inequity Aversion (Fehr and Schmidt (1999))

In a general version of the *inequity aversion* model, an agent derives the following utility:

$$V_i = U_i - \alpha_i \max_j (U_i - U_j, 0) - \beta_i \max_j (U_j - U_i, 0)$$

where U represents payoffs from the game and $0 \leq \alpha_i < \beta_i < 1$.[16] Hence, the type space is $\Theta^{IA} = \{(\alpha, \beta) : 0 \leq \alpha < \beta < 1\}$. We now define Prisoner's Dilemma games and population equilibrium for the inequity aversion model, as we did for the previous models.

Definition: A *Prisoner's Dilemma IA-Game* is a pair $< \Gamma, F >$ where Γ is a standard Prisoner's Dilemma game and F is a continuous distribution on Θ^{IA}.

[15] The Selten score is a measure of the predictive success of a model. It is the proportion of data consistent with the model minus the probability that any random data is consistent with that model. As both the switching cost model and the magical thinking model are consistent with the observed data, then the more restrictive model has a higher Selten score.

[16] An agent with utility V_i cares about his own payoff but is also averse to inequity. The parameters α_i and β_i specify how much aversion he feels towards unequal payoffs: α_i when the agent himself receives the higher payoff and β_i when the other agent receives the higher payoff. The condition $\alpha_i < \beta_i$ states that inequity bothers the agent more when he is receiving the lesser payoff. The requirement that $\alpha_i, \beta_i < 1$ implies that the agent is not willing to sacrifice his own payoff in order to achieve a more equal outcome. That is, holding the other player's payoff fixed, the agent aims to maximize his own payoff.

Definition: In a Prisoner's Dilemma IA-game $< \Gamma, F >$, an *IA-population equilibrium* is a measurable map $a : \Theta^{IA} \rightarrow \{Coop, Defect\}$ such that for every $\lambda = (\alpha, \beta) \in \Theta^{IA}$,

$$a(\lambda) \in \underset{a \in \{Coop, Defect\}}{\text{argmax}} \quad pU_1(a, Coop) + (1-p)U_1(a, Defect) - 6p\alpha\mathbb{1}_{a=Defect} - 6(1-p)\beta\mathbb{1}_{a=Coop}$$

where $p := F(\{\lambda \in \Theta^{IA} : a(\lambda) = Coop\})$.

For a fixed population F, we study IA-population equilibria of the Prisoner's Dilemma IA-games $< \Gamma_x, F >$. Unlike before, an IA-population equilibrium may not be unique.

Proposition 3.5 *For each Prisoner's Dilemma IA-game $< \Gamma_x, F >$ where F has full support and $x \leq 6$, there exists an IA-population equilibrium. However, uniqueness is not guaranteed and there is a Prisoner's Dilemma IA-game with multiple IA-population equilibria.*

For an IA-population equilibrium, we refer to p_x as the corresponding *inequity aversion cooperation rate*. While, for a given F, there may be multiple IA-population equilibria and thus multiple cooperation rates, the following proposition provides a necessary condition that all of these cooperation rates must satisfy.

Proposition 3.6 *Let p_x be the inequity aversion cooperation rate for an equilibrium of a Prisoner's Dilemma IA-game $< \Gamma_x, F >$ where F has full support and $x \in \{3, 4, 5, 6\}$. Then, either of the following must hold:*

- $p_x = 0$ *(trivial equilibrium)*
- $p_x > \dfrac{7}{6+x}$ *(non-trivial equilibrium)*

Notice that the necessary conditions derived in the above proposition are of a different sort. In the switching cost and magical thinking models, the derived conditions were across different p_x whereas here the conditions must be satisfied for each individual p_x.

Table 4 displays the cooperation rate conditions for each value of x. Generally, the above conditions for the inequity aversion model are not met by Charness et al.'s (2016) data.

x	Observed Coop Rate	Inequity Aversion Conditions
3	23.46%	$p_3 = 0\%$ or $p_3 > 77.78\%$
4	33.67%	$p_4 = 0\%$ or $p_4 > 70.00\%$
5	50.60%	$p_5 = 0\%$ or $p_5 > 63.64\%$
6	60.47%	$p_6 = 0\%$ or $p_6 > 58.33\%$

Table 4: Cooperation rates from Charness et al. (2016) and conditions derived from Proposition 3.6

Remark: The multiplicity of IA-population equilibria poses an obstacle to performing an analysis of cooperation rates across different x's as was done for the other models. Moreover, such an analysis would only serve to further restrict this model's predictions and therefore they would remain unsatisfied.

3.2.4 Fairness (Rabin (1993))

In order to apply the *fairness* model to the Prisoner's Dilemma game, we follow Daley and Sadowski's (2017) formulation and perform all possible algebraic reductions. Each agent has the following utility:

$$V_i(a_i, a_j) = U_i(a_i, a_j) + \lambda_i K(a_i, a_j), \text{ where } \lambda_i \geq 0 \text{ and } K(a_i, a_j) = \begin{cases} 3/4 & \text{if } a_i = a_j \\ -1/4 & \text{if } a_i \neq a_j. \end{cases}$$

Each agent's coefficient λ_i is therefore his type and the type space is $\Theta^{FAIR} = \mathbb{R}_+$. The kindness function K creates a coordination motivation: it is good for an agent to cooperate when his opponent cooperates and to defect when his opponent defects. This reflects an underlying motivation that an agent would like to be kind when his opponent is kind and mean when his opponent is mean.

We now define Prisoner's Dilemma games and population equilibrium for the fairness model, as we did for the previous models.

Definition: A *Prisoner's Dilemma FAIR-Game* is a pair $< \Gamma, F >$ where Γ is a standard Prisoner's Dilemma game and F is a continuous distribution on Θ^{FAIR}.

Definition: In a Prisoner's Dilemma FAIR-game $< \Gamma, F >$, a *$FAIR$-population equilibrium* is a measurable map $a : \Theta^{FAIR} \to \{Coop, Defect\}$ such that for every $\lambda \in \Theta^{FAIR}$,

$$a(\lambda) \in \underset{a \in \{Coop, Defect\}}{\text{argmax}} \; pU_1(a, Coop) + (1-p)U_1(a, Defect) + \left(\frac{3}{4} - p\right)\lambda \mathbb{1}_{a=Defect} + \left(p - \frac{1}{4}\right)\lambda \mathbb{1}_{a=Coop}$$

where $p := F(\{\lambda \in \Theta^{FAIR} : a(\lambda) = Coop\})$.

For a fixed population F, we study FAIR-population equilibria of the Prisoner's Dilemma FAIR-games $<\Gamma_x, F>$. As in inequity aversion, a FAIR-population equilibrium may not be unique.

Proposition 3.7 *For each Prisoner's Dilemma $FAIR$-game $<\Gamma_x, F>$ where F has full support and $x \leq 6$, there exists a FAIR-population equilibrium. However, uniqueness is not guaranteed and there is a Prisoner's Dilemma FAIR-game with infinitely many FAIR-population equilibria.*

To understand this proposition, notice that there is always a FAIR-population equilibrium where all players defect because both material incentives and the kindness (meanness) motivation lead an agent to defect when all other players defect. On the other hand, the coordination motivation (created by the kindness function) often gives rise to a cooperative equilibrium where a majority of agents cooperate. In fact, there are distributions of agents for which there are infinitely many cooperative equilibria.

For a FAIR-population equilibrium, we refer to p_x as the corresponding *fairness cooperation rate*. As noted, for a given F, there may be multiple FAIR-population equilibria and thus multiple cooperation rates. The following proposition provides a necessary condition which all of the cooperation rates must satisfy.

Proposition 3.8 *Let p_x be the fairness cooperation rate for an equilibrium of a Prisoner's Dilemma FAIR-game $<\Gamma_x, F>$ where F has full support and $x \in \{3,4,5,6\}$. Then, either of the following must hold:*

- $p_x = 0$ *(defect equilibrium)*
- $p_x > \frac{1}{2}$ *(cooperative equilibrium)*

Therefore, the fairness model requires that all cooperation rates are either 0 or above 50%, and this is inconsistent with Charness et al.'s (2016) data.

x	Observed Coop Rate	Fairness Conditions
3	23.46%	$p_3 = 0\%$ or $p_3 > 50\%$
4	33.67%	$p_4 = 0\%$ or $p_4 > 50\%$
5	50.60%	$p_5 = 0\%$ or $p_5 > 50\%$
6	60.47%	$p_6 = 0\%$ or $p_6 > 50\%$

Table 5: Cooperation rates from Charness et al. (2016) and conditions derived from Proposition 3.8

3.3 Finitely Repeated Prisoner's Dilemma Games

Up until this point, we have focused exclusively on one-shot Prisoner's Dilemma games. Classically, Kreps et al. (1982) demonstrate that cooperation can be sustained in a finitely repeated Prisoner's Dilemma game when there is a small probability that a player is a committed Tit-for-Tat type.[17] Interestingly, in all sequential equilbria (Kreps and Wilson (1982)), rational players who have full access to all strategies may then choose to imitate the Tit-for-Tat player in order to develop a reputation that he is actually a Tit-for-Tat player. His opponent, not knowing the player's type, may end up cooperating as well, rather than testing him. This cooperation crucially depends on the game being repeated. In a one-shot setting, there is no possibility of developing a reputation and thus no rational type would cooperate. Thus, the introduction of behavioral types is a method that can sustain cooperation *only* in repeated settings but not in one-shot settings.

As mentioned earlier, Lipman and Wang (2000) study a model in which an agent incurs a switching cost every time he alters his strategy from one stage to another. They find that cooperation in the Prisoner's Dilemma game depends upon a sensitive trade-off, since the switching costs reduce both the incentives to defect from cooperation and the incentives to punish others' defections. As in the commitment models above, Lipman and Wang's (2000) model can sustain cooperation *only* in repeated settings.

On the other hand, our switching cost model can sustain cooperation *even* in one-shot settings. Thus, without any additional features (such as behavioral types), it can also sustain cooperation in a finitely repeated setting, since the repetition of the one-shot equilibrium is an equilibrium of the repeated game. Committed behavioral types can also be added, if one wants a model with: (i) cooperation in both one-shot and repeated settings, and (ii) more cooperation in repeated settings than in one-shot settings (as in Cooper et al.'s (1996) experimental findings).

[17] A Tit-for-Tat player plays *Coop* in the first stage. In every other stage, he imitates the other player's previous play. This strategy is called Tit-for-Tat because it has the sensible logic: "If you cooperated yesterday, then so will I today, but if you defected, then I am going to punish you by defecting today."

4 Concluding Remarks

We propose and axiomatically characterize a status quo-based choice model where an agent incurs a cost when switching from a status quo to another alternative. In a choice problem with a status quo, the agent chooses the alternatives with the highest utility net of this switching cost. This model accommodates status quo bias and general reference effects.

We apply the model to a particular class of Prisoner's Dilemma games, and establish necessary and sufficient conditions for a cooperation rate in these games to be generated by our model. Using Charness et al.'s (2016) experimental data, we verify that these conditions hold. We then compare our model's theoretical predictions with those of other theories for this class of Prisoner's Dilemma games and find that our theory provides the tightest predictions that are consistent with Charness et al.'s (2016) data. While our results are tailored to Charness et al.'s (2016) domain, they also hold true in other settings.[18]

It has been suggested that status quos can be paternalistically effective in influencing people's decisions in a variety of areas from healthcare to investment (Samuelson and Zeckhauser (1988), Madrian and Shea (2001), Choi et al. (2004), and Thaler and Sunstein (2008)). Our model provides a more nuanced view of a status quo's impact. First, status quos might not only rule alternatives out, as described in the literature, but may also steer agents' behavior in other directions. Second, if switching costs provide a basis for the status quo bias phenomenon, then one needs to be careful when evaluating the welfare implications of paternalistic policies. This is because, for a proper accounting, the benefits that come from superior choices need to be weighed against the switching costs incurred by the agents.[19] We hope that this viewpoint proves useful in such policy work.

Finally, a question that naturally arises is whether the switching cost model can also be employed in extensive form games in general. One approach would be to consider the normal form of such games (that is, the one-shot game where players choose strategies for the extensive form game) and then to define switching costs between strategies (rather than between actions). This strategy-based switching cost should incorporate the dynamic structure of the game, for example, one natural switching cost function would be the discounted sum of switching costs from each node.

[18] For example, for Proposition 3.2, if we had data for a continuous range of x, then the conditions would become: p_x is strictly increasing in x; $(6-x)\cdot p_x < (6-y)\cdot p_y$ for all $x < y < 6$; and $0 < p_x < 1$.

[19] Arad and Rubinstein (2018) conduct a survey regarding paternalistic policies. While a majority of respondents support such policies, a substantial proportion opposes them and switching costs could provide a basis for such a rejection.

References

[1] AKERLOF, G. A., AND KRANTON, R. E. Economics and Identity. *The Quarterly Journal of Economics 115* (2005), 715–753.

[2] APESTEGUIA, J., AND BALLESTER, M. A Theory of Reference-Dependent Behavior. *Economic Theory 40* (2009), 427–455.

[3] ARAD, A., AND RUBINSTEIN, A. The People's Perspective on Libertarian-Paternalistic Policies. *Journal of Law and Economics* (2018), forthcoming.

[4] AUSUBEL, L. M. The Failure of Competition in the Credit Card Market. *The American Economic Review 81* (1991), 50–81.

[5] BEGGS, A., AND KLEMPERER, P. Multi-Period Competition with Switching Costs. *Econometrica 60* (1992), 651–666.

[6] BOSSERT, W., AND SPRUMONT, Y. Efficient and Non-Deteriorating Choice. *Mathematical Social Sciences 45* (2003), 131–142.

[7] CARLSSON, F., AND LÖFGREN, Å. Airline Choice, Switching Costs and Frequent Flyer Programmes. *Applied Economics 38* (2006), 1469–1475.

[8] CARUANA, G., AND EINAV, L. A Theory of Endogenous Commitment. *The Review of Economic Studies 75* (2008), 99–116.

[9] CHARNESS, G., RIGOTTI, L., AND RUSTICHINI, A. Social Surplus Determines Cooperation Rates in the One-Shot Prisoner's Dilemma. *Games and Economic Behavior 100* (2016), 113–124.

[10] CHEVALIER, J. A., AND SCHARFSTEIN, D. S. Capital-Market Imperfections and Countercyclical Markups: Theory and Evidence. *The American Economic Review 86* (1996), 703–725.

[11] CHOI, J. J., LAIBSON, D., MADRIAN, B., AND METRICK, A. For Better or For Worse: Default Effects and 401(k) Savings Behavior. *Perspectives on the Economics of Aging* (2004), 81–121.

[12] Cooper, R., DeJong, D. V., Forsythe, R., and Ross, T. W. Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games. *Games and Economic Behavior 12* (1996), 187–218.

[13] Daley, B., and Sadowski, P. Magical Thinking: A Representation Result. *Theoretical Economics 12* (2017), 909–956.

[14] Echenique, F., Lee, S., and Shum, M. The Money Pump as a Measure of Revealed Preference Violations. *Journal of Political Economy 119* (2011), 1201–1223.

[15] Farquhar, P. H., and Pratkanis, A. R. A Brief History of Research on Phantom Alternatives: Evidence for Seven Empirical Generalizations About Phantoms. *Basic and Applied Social Psychology 13* (1992), 103–122.

[16] Farquhar, P. H., and Pratkanis, A. R. Decision Structuring with Phantom Alternatives. *Management Science 39* (1993), 1214–1226.

[17] Farrell, J., and Shapiro, C. Optimal Contracts with Lock-In. *The American Economic Review 79* (1989), 51–68.

[18] Fehr, E., and Schmidt, K. M. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics 114* (1999), 817–868.

[19] Gazzaniga, M. S. The Ethical Brain. *New York: Dana Press* (2005).

[20] Guney, B., and Richter, M. An Experiment on Aspiration-Based Choice. *Journal of Economic Behavior and Organization 119* (2015), 512–526.

[21] Guney, B., Richter, M., and Tsur, M. Aspiration-Based Choice. *Journal of Economic Theory 176* (2018), 935–956.

[22] Highhouse, S. Context-Dependent Selection: The Effects of Decoy and Phantom Job Candidates. *Organizational Behavior and Human Decision Processes 65* (1996), 68–76.

[23] Jaffray, J. Y. Semicontinuous Extension of a Partial Order. *Journal of Mathematical Economics 2* (1975), 395–406.

[24] KALAI, E., AND SMORODINSKY, M. Other Solutions to Nash's Bargaining Problem. *Econometrica 43* (1975), 513–518.

[25] KLEMPERER, P. Markets with Consumer Switching Costs. *The Quarterly Journal of Economics 102* (1987), 375–394.

[26] KLEMPERER, P. Competition When Consumers Have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade. *The Review of Economic Studies 62* (1995), 515–539.

[27] KNETSCH, J. L. The Endowment Effect and Evidence of Nonreversible Indifference Curves. *The American Economic Review 79* (1989), 1277–1284.

[28] KREPS, D. M., MILGROM, P., ROBERTS, J., AND WILSON, R. Rational Cooperation in the Finitely Repeated Prisoners' Dilemma. *Journal of Economic Theory 27* (1982), 245–252.

[29] KREPS, D. M.AND WILSON, R. Sequential Equilibria. *Econometrica 50* (1982), 863–894.

[30] LEVITT, S., AND LIST, J. A. What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World? *Journal of Economic Perspectives 21* (2007), 153–174.

[31] LIBICH, J. An Explicit Inflation Target as a Commitment Device. *Journal of Macroeconomics 30* (2008), 43–68.

[32] LIPMAN, B. L., AND WANG, R. Switching Costs in Frequently Repeated Games. *Journal of Economic Theory 93* (2000), 149–190.

[33] LIPMAN, B. L., AND WANG, R. Switching Costs in Infinitely Repeated Games. *Games and Economic Behavior 66* (2009), 292–314.

[34] MADRIAN, B., AND SHEA, D. The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior. *The Quarterly Journal of Economics 116* (2001), 1149–1187.

[35] MASATLIOGLU, Y., AND OK, E. Rational Choice with Status Quo Bias. *Journal of Economic Theory 121* (2005), 1–29.

[36] MASATLIOGLU, Y., AND OK, E. A Canonical Model of Choice with Initial Endowments. *The Review of Economic Studies 81* (2014), 851–883.

[37] MASATLIOGLU, Y., AND ULER, N. Understanding the Reference Effect. *Games and Economic Behavior 82* (2013), 403–423.

[38] MUNRO, A., AND SUGDEN, R. On the Theory of Reference-Dependent Preferences. *Journal of Economic Behavior and Organization 50* (2003), 407–428.

[39] OK, E., ORTOLEVA, P., AND RIELLA, G. Revealed (P)reference Theory. *The American Economic Review 105* (2014), 299–321.

[40] ORTOLEVA, P. Status Quo Bias, Multiple Priors and Uncertainty Aversion. *Games and Economic Behavior 69* (2010), 411–424.

[41] PETTIBONE, J. C., AND WEDELL, D. H. Examining Models of Nondominated Decoy Effects Across Judgment and Choice. *Organizational Behavior and Human Decision Processes 81* (2000), 300–328.

[42] RABIN, M. Incorporating Fairness into Game Theory and Economics. *The American Economic Review 83* (1993), 1281–1302.

[43] RADER, T. On the Existence of Utility Functions to Represent Preferences. *The Review of Economic Studies 30* (1963), 229–232.

[44] RIELLA, G., AND TEPER, R. Probabilistic Dominance and Status Quo Bias. *Games and Economic Behavior 87* (2014), 288–304.

[45] RUBINSTEIN, A., AND ZHOU, L. Choice Problems with a 'Reference' Point. *Mathematical Social Sciences 37* (1999), 205–209.

[46] SAGI, J. S. Anchored Preference Relations. *Journal of Economic Theory 130* (2006), 283–295.

[47] SALANT, Y., AND RUBINSTEIN, A. (A, f): Choice with Frames. *The Review of Economic Studies 75* (2008), 1287–1296.

[48] SAMUELSON, W., AND ZECKHAUSER, R. Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty 1* (1988), 7–59.

[49] SELTEN, R. Properties of a Measure of Predictive Success. *Mathematical Social Sciences 21* (1991), 153–167.

[50] SHI, M., CHIANG, J., AND RHEE, B. D. Price Competition with Reduced Consumer Switching Costs: The Case of Wireless Number Portability in the Cellular Phone Industry. *Management Science 52* (2006), 27–38.

[51] SUGDEN, R. Alternatives to the Neo-classical Theory of Choice. *Valuing Environmental Preferences, Oxford University Press.* (1999).

[52] THALER, R., AND SUNSTEIN, C. Nudge: Improving Decisions About Health, Wealth, and Happiness. *Yale University Press* (2008).

[53] TVERSKY, A., AND KAHNEMAN, D. Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics 106* (1991), 1039–1061.

[54] VON WEIZSÄCKER, C. C. The Costs of Substitution. *Econometrica 52* (1984), 1085–1116.

Appendix I

(A) Proof of Section 2 Results

Proof of Theorem 2.1

[**Sufficiency**] Take any choice correspondence c that satisfies Axioms WARP, SQB, and UHC. Define the reference-free preference $\succeq$ as follows: $a \succeq b$ if $a \in c(ab, \diamond)$. The preference relation $\succeq$ is continuous by Axiom UHC (ii) and complete by Axiom WARP such that $c(S, \diamond) = \max_{s \in S}(S, \succeq)$ for any $(S, \diamond) \in \mathcal{C}_\diamond(X)$.

Now, define $\succeq^*$ as follows:

$(x, y) \succeq^* (z, t)$ if either $y = t$ and $x \in c(xz, y)$, or there exists $e \in X$ such that (1) $e \in c(ez, t)$, (2) $y \succeq e$, and (3) $x \in c(xy, y)$.

Notice that $\succeq^*$ is reflexive and moreover $(a, a) \succeq^* (a, b)$ for any $a, b \in X$.[20]

Claim 0: For any $x, y, z \in X$, if $(x, y) \succeq^* (z, y)$, then $x \in c(xz, y)$.

Proof of Claim 0: The preference relation $\succeq^*$ is defined according to two conditions. We now show that in the case above, the second condition implies the first, that is, if there exists an e such that (1) $e \in c(ez, y)$, (2) $y \succeq e$, and (3) $x \in c(xy, y)$, then $x \in c(xz, y)$. By definition, (2) states that $y \in c(ey, \diamond)$. By Axiom SQB(i), $y \in c(ey, y)$. Suppose that $x \notin c(exyz, y)$. Then, $c(exyz, y) \cap \{x, y\} \neq c(xy, y)$, so by WARP it must be that $y \notin c(exyz, y)$. But, then $c(exyz, y) \cap \{e, y\} \neq c(ey, y)$ and again by WARP, it must be that $e \notin c(exyz, y)$. Thus, $z = c(exyz, y)$ and WARP implies $z = c(ez, y)$, a contradiction. Therefore, $x \in c(exyz, y)$ and by WARP, $x \in c(xz, y)$.

Claim 1: For any $x, y \in X$, $(x, x) \succeq^* (y, y)$ if and only if $x \succeq y$.

Proof of Claim 1: To prove the if direction, let e in the definition of $\succeq^*$ be equal to y. To prove the only if direction, suppose $(x, x) \succeq^* (y, y)$ for some $x, y \in X$. Then, by definition of $\succeq^*$, there must exist $e \in X$ such that $e \in c(ey, y)$ and $x \succeq e$. If $e \notin c(ey, \diamond)$, then $y = c(ey, \diamond)$ and axiom SQB(ii) implies that $y = c(ey, y)$, contradicting the fact that $e \in c(ey, y)$. Therefore, $e \in c(ey, \diamond)$. Since $\succeq$ represents $c(\cdot, \diamond)$, we have $e \succeq y$, which in turn implies $x \succeq y$.

[20] Simply choose $e = a$ in the definition of $\succeq^*$.

Claim 2: $c(S,x) = \max_{s\in S}((s,x), \succeq^*)$

Proof of Claim 2: Let $y \in c(S,x)$. Axiom WARP implies $y \in c(yz,x)$ for any $z \in S$. Then, by definition of $\succeq^*$, we obtain $(y,x) \succeq^* (z,x)$ for any $z \in S$. To show the other inclusion, suppose $s \in S$ such that $(s,x) \succeq^* (t,x)$ for any $t \in S$ but $s \notin c(S,x)$. Since $c(S,x) \neq \emptyset$, there must exist $r \in c(S,x)$. Axiom WARP implies $r \in c(rs,x)$. Since (s,x) is the maximizer of $\succeq^*$ on S, we have $(s,x) \succeq^* (r,x)$. By Claim 0, $s \in c(rs,x)$. Therefore $s \in c(rs,x)$ and so $\{r,s\} = c(rs,x)$. As $r \in c(S,x)$, Axiom WARP states that $s \in c(S,x)$, which is a contradiction.

Claim 3: $\succeq^*$ is transitive.

Proof of Claim 3: Suppose $(x,y) \succeq^* (z,t)$ and $(z,t) \succeq^* (m,n)$ for some $x,y,z,t,m,n \in X$, and we need to show $(x,y) \succeq^* (m,n)$. There are four possible cases:

Case 1: $y = t = n$

By Claim 0, $x \in c(xz,y)$ and $z \in c(zm,y)$. By standard arguments, since $c(\cdot,y)$ satisfies WARP, it is the case that $x \in c(xm,y)$. Therefore, $(x,y) \succeq^* (m,y)$.

Case 2: $y = t$ and $t \neq n$

By Claim 0, $x \in c(xz,y)$. By the definition of $\succeq^*$, there exists e such that $e \in c(em,n)$, $y \succeq e$, and $z \in c(zy,y)$. By Case 1, $x \in c(xy,y)$. By the definition of $\succeq^*$, we have $(x,y) \succeq^* (m,n)$.

Case 3: $t = n$ and $y \neq t$

Again, by Claim 0, $z \in c(zm,t)$ and by the definition of $\succeq^*$, there exists an e such that $e \in c(ez,t)$, $y \succeq e$, and $x \in c(xy,y)$. By Case 1, $e \in c(em,t)$.[21] The definition of $\succeq^*$ then states $(x,y) \succeq^* (m,t)$.

Case 4: $t \neq n$ and $y \neq t$

$(x,y) \succeq^* (z,t)$ implies that there exists $e \in X$ such that $e \in c(ez,t)$, $y \succeq e$, and $x \in c(xy,y)$. Similarly, $(z,t) \succeq^* (m,n)$ guarantees that there exists $f \in X$ such that $f \in c(mf,n)$, $t \succeq f$, and $z \in c(zt,t)$. Then, we obtain

$$(y,y) \succeq^* (e,e) \succeq^* (e,t) \succeq^* (z,t) \succeq^* (t,t) \succeq^* (f,f).$$

Applying transitivity of $\succeq^*$ to case 3 type of pairs reduces the above to

$$(y,y) \succeq^* (e,e) \succeq^* (t,t) \succeq^* (f,f)$$

Claim 1 together with the transitivity of $\succeq$ gives $y \succeq f$. Then, by the definition of $\succeq^*$, we conclude that $(x,y) \succeq^* (m,n)$.

[21] Case 1 is applied to $(e,t) \succeq^* (z,t)$ and $(z,t) \succeq^* (m,t)$.

Claim 4: $\succeq^*$ is upper semi-continuous.

Proof of Claim 4: Suppose $(x_n, y_n) \succeq^* (z,t)$ for all n, where $x_n, y_n, x, y, z, t \in X$ such that $x_n \to x$ and $y_n \to y$. We need to show $(x,y) \succeq^* (z,t)$.

Suppose $y_n = t$ for infinitely many n. Then, by Claim 0, it is the case that $x_n \in c(x_n z, t)$ for infinitely many n. Axiom UHC (iii) implies $x \in c(xz, t)$. Moreover, by the definition of a limit, $y = t$. Thus, $(x,y) \succeq^* (z,t)$.

Suppose $y_n \neq t$ for infinitely many n. Then, there exist infinitely many $e_n \in X$ such that $e_n \in c(e_n z, t)$, $y_n \succeq e_n$, and $x_n \in c(x_n y_n, y_n)$. Since X is compact, there exists a convergent subsequence e_{n_k} of e_n such that $e_{n_k} \to e$ for some $e \in X$. Axiom UHC (i) gives $e \in c(ez, t)$, $y \succeq e$, and $x \in c(xy, y)$. According to the definition of $\succeq^*$, we obtain $(x,y) \succeq^* (z,t)$.

Remember that $\succeq^*$ is reflexive, transitive, and upper semi-continuous while not necessarily complete over the entire space. Jaffray's (1975) result guarantees that $\succeq^*$ can be extended to a reflexive, transitive, upper semi-continuous, and complete relation denoted by $\succeq^*_E$.

Remark: As shown in Claim 2, given a fixed status quo, the original preference relation is complete and rationalizes choices. The extended relation respects this ranking and therefore it rationalizes choices, given a fixed status quo.

Rader's (1963) result ensures the existence of an upper semi-continuous utility function $\overline{U} : X \times X \to \mathbb{R}$ that represents $\succeq^*_E$. This, together with the remark above, gives

$$c(S,x) = \operatorname*{argmax}_{s \in S} \overline{U}(s,x) \quad \text{for any } (S,x) \in \mathcal{C}_{sq}(X).$$

Claim 1 and the fact that $\succeq^*_E$ respects ranking by $\succeq^*$ allows us to define $U : X \to \mathbb{R}$ by $U(x) := \overline{U}(x,x)$ for any $x \in X$ so that U rationalizes $c(\cdot, \diamond)$. We also define $D : X \times X \to \mathbb{R}_+$ by setting $D(s,x) = -\overline{U}(s,x) + \overline{U}(s,s)$ for any $s, x \in X$. Then,

$$c(S,x) = \operatorname*{argmax}_{s \in S} U(s) - D(s,x) \quad \text{for any } (S,x) \in \mathcal{C}_{sq}(X).$$

Notice first that $D(x, x) = 0$ for any $x \in X$. Second, D is lower semi-continuous since $\overline{U}$ is upper semi-continuous and U is continuous. Moreover, $D(s, x) \geq 0$ for any $x, s \in X$ because $(s, s) \succeq^* (s, x)$ and therefore $(s, s) \succeq_E^* (s, x)$ which in turn implies $\overline{U}(s, s) - \overline{U}(s, x) \geq 0$.

[**Necessity**] Take any continuous utility function $U : X \to \mathbb{R}$ and a lower semi-continuous function $D : X \times X \to \mathbb{R}_+$ with $D(x, x) = 0$ for any $x \in X$. Suppose also that $c(S, \diamond) = \underset{s \in S}{\operatorname{argmax}}\, U(s)$ for any $(S, \diamond) \in \mathcal{C}_\diamond(X)$ and $c(S, x) = \underset{s \in S}{\operatorname{argmax}}\, U(s) - D(s, x)$ for any $(S, x) \in \mathcal{C}_{sq}(X)$.

Axiom WARP is trivially satisfied across a fixed status quo position. To see Axiom SQB(i), take $y \in c(xy, \diamond)$. Then $U(y) \geq U(x)$ and $D(y, y) = 0 \leq D(x, y)$. Therefore $U(y) - D(y, y) \geq U(x) - D(x, y)$ $\Rightarrow y \in c(xy, y)$. To see Axiom SQB(ii), take $y = c(xy, \diamond)$. Then $U(y) > U(x)$ and $D(y, y) = 0 \leq D(x, y)$. Therefore $U(y) - D(y, y) > U(x) - D(x, y) \Rightarrow y = c(xy, y)$.

To show Axiom UHC (i), suppose $y_n \in c(y_n x_n, x_n)$ for some y_n, x_n, y, x such that $y_n \to y$ ad $x_n \to x$. By hypothesis, $U(y_n) - D(y_n, x_n) \geq U(x_n) - D(x_n, x_n)$ for all n. As $D(x_n, x_n) = 0$ for all n, the inequality reduces to $U(y_n) - D(y_n, x_n) \geq U(x_n)$ for all n. Since U is continuous and $-D$ is upper semi-continuous, taking the limsup of both sides gives $U(y) - D(y, x) \geq U(x)$ which is equivalent to $U(y) - D(y, x) \geq U(x) - D(x, x)$. This implies $y \in c(yx, x)$ as desired. To verify Axiom UHC (ii), it is enough to notice that U is continuous. Finally, to prove that Axiom UHC (iii) is also satisfied, assume $y_n \in c(y_n x, z)$ for some y_n, y, x, z such that $y_n \to y$. This means $U(y_n) - D(y_n, z) \geq U(x) - D(x, z)$ for all n. Taking limsup of both sides and the continuity of U together with the upper semi-continuity of $-D$ yields $U(y) - D(y, z) \geq U(x) - D(x, z)$. Thus, $y \in c(yx, z)$. □

(B) Proof of Section 3 Results

Lemma 3.0 In a Prisoner's Dilemma D-game $< \Gamma_x, F >$ with a population equilibrium a_x, an agent cooperates if and only if his status quo is *Coop* and $(6 - x) \cdot p_x + 1 < D_\theta(Defect, Coop)$.

Proof of Lemma 3.0

An agent with a status quo of $Defect$ will necessarily defect as switching to *Coop* would lower his utility and he would also incur a switching cost, two negative effects. On the other hand, if an agent has a status quo of *Coop*, he will stick with *Coop* if and only if the following holds:

$$\mathbb{E}\left[U_1(Defect, a_2)\right] - D_1(Defect, Coop) < \mathbb{E}\left[U_1(Coop, a_2)\right]$$

$$7 \cdot p_x + 2(1 - p_x) - D_1(Defect, Coop) < x \cdot p_x + 1 \cdot (1 - p_x)$$
$$(7 - x) \cdot p_x + (1 - p_x) < D_1(Defect, Coop)$$
$$(6 - x) \cdot p_x + 1 < D_1(Defect, Coop)$$

The agent will choose $Coop$ if the above holds strictly, will choose $Defect$ if the reverse holds strictly, and will be indifferent between $Coop$ and $Defect$ if there is an equality. Tie-breaking can be done in any fashion because the distribution is continuous, and so only for a measure 0 set of agents can the above hold with equality.

□

Proof of Proposition 3.1

[Existence] The cooperation rate p is an equilibrium if $p = 1 - F((6 - x) \cdot p + 1)$. Notice that when $p = 0$, the LHS is strictly less than the RHS because of the full support condition. On the other hand, when $p = 1$, the LHS is strictly higher than the RHS for the same reason. By continuity, there is some $\hat{p}$ which solves this condition.

[Uniqueness] Suppose not. Then, there are at least two different population equilibria with cooperation rates $p < q$. By Lemma 3.0, an agent will cooperate if his status quo is $Coop$ and his switching costs are strictly above the cooperation threshold, and will defect if his switching cost is strictly below the cooperation threshold. Therefore, in equilibrium, $p = 1 - F((6 - x) \cdot p + 1)$ and $q = 1 - F((6 - x) \cdot q + 1)$. So, $p < q$ implies $F((6 - x) \cdot p + 1) > F((6 - x) \cdot q + 1)$. Furthermore, F is strictly increasing because it is a distribution function with full support, so it must be that $(6 - x) \cdot p + 1 > (6 - x) \cdot q + 1 \Rightarrow p > q$ if $x < 6$. If $x = 6$, then $p = 1 - F((6 - x) \cdot p + 1) = 1 - F(1) = 1 - F((6 - x) \cdot q + 1) = q$. Either way, we arrive at a contradiction, so there cannot be two distinct cooperation rates p, q. □

Proof of Proposition 3.2

[**Necessity**] For the non-triviality condition, notice that in equilibrium $p_x = 1 - F((6 - x) \cdot p_x + 1)$ and since F has full support, it must be that $0 < p_x < 1$.

For the monotonicity condition, take $x < y \leq 6$. Due to non-triviality, it must be that $0 < p_x < 1$ and similarly $0 < p_y < 1$. Now, suppose that $p_x \geq p_y$. Then, $(6 - x) \cdot p_x + 1 > (6 - y) \cdot p_y + 1$ and because F has full support, it is the case that $p_x = 1 - F((6 - x) \cdot p_x + 1) < 1 - F((6 - y) \cdot p_y + 1) = p_y$, which is a contradiction.

For the threshold condition, notice that we have $p_x = F(\theta : s_\theta = Coop,\ (6-x) \cdot p_x + 1 < D_\theta)$ by Lemma 3.0. By the monotonicity condition proved above, $p_{x-1} < p_x$ and so the threshold $(6-x) \cdot p_x + 1$ has to be strictly decreasing in x. Rearranging yields the identity.

[**Sufficiency**] For any given p_x that satisfies the above two properties, we must construct a distribution F with full support which generates cooperation rate p_x. For now, we momentarily assume that all agents have a $Coop$ status quo, so that F is simply a distribution over switching costs. Lemma 3.0 established that a type θ cooperates in Γ_x if $(6-x)p_x + 1 < D_\theta(Defect, Coop)$. As the distribution is continuous with full support, switching cost threshold is $d_x = (6-x)p_x + 1$. The distribution F must then satisfy that $1 - F(d_x) = p_x$, or $F(d_x) = 1 - p_x$. That is, there are four given cooperation rates p_3, p_4, p_5, p_6 and four induced switching cost thresholds d_3, d_4, d_5, d_6, and F is defined so that $F(d_x) = 1 - p_x$ for each x. But, it needs to be verified that F is indeed a distribution, i.e. it is increasing, non-negative, and less than 1. Condition 1 establishes that $p_3 < p_4 < p_5 < p_6$ and Condition 2 establishes that $(6-x)p_x < (7-x)p_{x-1} \Rightarrow (6-x)p_x + 1 < (6-(x-1))p_{x-1} + 1 \Rightarrow d_x < d_{x-1}$. Therefore, $d_3 > d_4 > d_5 > d_6$, establishing that F is increasing. Because $0 < p_x < 1$, so is $0 < 1 - p_x < 1$ and therefore, $0 < F < 1$ for the defined x-values. Now, we can take any F with full support over switching costs which satisfies these point restrictions. Then, we can take a full support F so that a fraction of agents with switching costs beneath the lowest threshold, d_3, are reassigned to have $Defect$ as a status quo. Those agents will always defect, regardless of whether their status quo is $Defect$ or $Coop$ and so can be safely reassigned without disturbing the cooperation rate so as to guarantee the full support of F. □

Proof of Proposition 3.3

In a Prisoner's Dilemma MT-Game $< \Gamma_x, F >$, given that other players are cooperating with probability p, an agent with type λ chooses Coop if and only if

$$\lambda \cdot x + (1-\lambda) \cdot (xp + 1 \cdot (1-p)) \geq 2\lambda + (1-\lambda) \cdot (7p + 2 \cdot (1-p)) \Leftrightarrow$$
$$\lambda \cdot (x-2) + (1-\lambda) \cdot ((x-7)p - (1-p)) \geq 0$$

The second equation is a convex combination of the positive term $x - 2$ and the negative term $((x - 7)p - (1-p)) = (x-6)p - 1$. Therefore, for any given cooperation rate p, there will be a unique threshold $\hat{\lambda}$ such that types with λ above this threshold will cooperate while types below this threshold will defect.

This threshold can be calculated by solving:

$$\hat{\lambda} \cdot (x-2) + (1-\hat{\lambda}) \cdot ((x-6)p - 1) = 0 \tag{*}$$

Also, as only agents above this threshold cooperate, the following must also hold:

$$p = 1 - F(\hat{\lambda}) \tag{**}$$

Suppose $x < 6$. If we solve for p in $(*)$, we see that p is increasing in $\hat{\lambda}$. If we solve for p in $(**)$, we see that p is decreasing in $\hat{\lambda}$. Since an increasing and decreasing function can intersect at most once, there is at most a unique cooperation rate p and threshold $\hat{\lambda}$ which satisfies the equilibrium conditions. Moreover, that there is an intersection can be gleaned from the following fact:

$$\text{If } p = 0 \text{ then } \overbrace{\lambda = 1/(x-1)}^{(*)} < \overbrace{\lambda = \infty}^{(**)}, \text{ and if } p = 1 \text{ then } \overbrace{\lambda = (7-x)/5}^{(*)} > \overbrace{\lambda = 0}^{(**)}.$$

Therefore, when $x < 6$, an MT-population equilibrium exists and is unique.

Finally, for the case that $x = 6$, conditon (*) is $\hat{\lambda} \cdot 4 + (1-\hat{\lambda}) \cdot (-1) = 0$ which implies $\hat{\lambda} = 1/5$ and by (**), p is uniquely defined as $p = 1 - F(1/5)$. Therefore, when $x = 6$, an MT-population equilibrium exists and is unique. □

Proof of Proposition 3.4

[**Necessity**]

1. Equation (*) above defines $\hat{\lambda}(x-2) + (1-\hat{\lambda})((x-6)p_x - 1) = 0$ which can then be re-expressed as $\hat{\lambda} = \dfrac{(6-x)p_x + 1}{(6-x)p_x + x - 1} = 1 - \dfrac{x-2}{(6-x)p_x + x - 1}$. Notice that if x strictly increases and p_x weakly decreases, then $\hat{\lambda}$ strictly decreases. But, if the threshold $\hat{\lambda}$ strictly decreases, then the cooperation rate p_x must strictly increase (because $1 - F(\hat{\lambda}) = p_x$), which is a contradiction. Therefore, as x strictly increases, it must be that p_x strictly increases.

2. For the cooperation rates to have the relationship $p_{x-1} < p_x$ it must be that the threshold $\hat{\lambda}$ is strictly decreasing in x. That is, $1 - \dfrac{x-3}{(7-x)p_{x-1} + x - 2} > 1 - \dfrac{x-2}{(6-x)p_x + x - 1}$. Therefore, $(x-3)((6-x)p_x + x - 1) < (x-2)((7-x)p_{x-1} + x - 2) \Rightarrow (x-3)(6-x)p_x < (x-2)(7-x)p_{x-1} + 1$.

Rearranging yields $p_x < \dfrac{(x-2)(7-x)}{(x-3)(6-x)}p_{x-1} + \dfrac{1}{(x-3)(6-x)}$. Notice, the only x for which p_x and p_{x-1} can be related are $x = 4, 5, 6$, but the condition holds vacuously for $x = 6$. Therefore, we are left with the condition for $x = 4, 5$.

3. An agent cooperates if and only if $\lambda \cdot (x-2) + (1-\lambda) \cdot ((x-7)p_x - (1-p_x) \geq 0$. This condition is a convex combination of two terms, one positive and one negative. Therefore, for any fixed $x \leq 6$, it always holds if λ is sufficiently close to 1 and never holds if λ is sufficiently close to 0. Therefore, given full support, there will be a non-zero measure of cooperators and a non-zero measure of defectors.

[**Sufficiency**] For any given cooperation rate p_x that satisfies the properties above, we must construct a distribution F with full support which generates cooperation rate p_x. From the previous analysis, we have established that a type λ cooperates if and only if $\lambda(x-2) + (1-\lambda)((x-7)p_x - (1-p_x)) \geq 0$. As the distribution is continuous with full support, the threshold λ_x is defined as $\lambda_x(x-2) + (1-\lambda_x)((x-7)p_x - (1-p_x)) = 0$. That is, agents with magical thinking levels above λ_x cooperate and those with magical thinking levels below defect. The distribution F must then satisfy $1 - F(\lambda_x) = p_x$ or $F(\lambda_x) = 1 - p_x$. That is, there are four given cooperation rates p_3, p_4, p_5, p_6 and four induced thresholds $\lambda_3, \lambda_4, \lambda_5, \lambda_6$, and F is defined so that $F(\lambda_x) = 1 - p_x$ for each x. But, it needs to be verified that F is indeed a distribution, i.e. it is increasing, non-negative, and less than 1. Condition 1 establishes that $p_3 < p_4 < p_5 < p_6$ and Condition 2 establishes that $\lambda_5 < \lambda_4 < \lambda_3$. To see that $\lambda_6 < \lambda_5$, notice that $\lambda_6 = 1/5$ and that $p_4 > 0 \Rightarrow \lambda_5 > 1/4$. This establishes that F is increasing. Since $0 < p_x < 1$, so is $0 < 1 - p_x < 1$ and therefore, $0 < F < 1$ for the defined x-values. Now, we can take any F with full support over magical thinking levels which satisfies these point restrictions.

(ii) From Proposition 3.2, the switching cost threshold condition is $p_x < \left(\dfrac{7-x}{6-x}\right) \cdot p_{x-1}$. This upper bound can be rewritten as $\dfrac{(x-3)(7-x)}{(x-3)(6-x)}p_{x-1} < \dfrac{(x-2)(7-x)}{(x-3)(6-x)}p_{x-1} + \dfrac{1}{(x-3)(6-x)}$. Therefore, Condition 2 of Proposition 3.2 is tighter than Condition 2 of Proposition 3.4. □

Proof of Proposition 3.5

[Existence] Recall that

$$V_i = U_i - \alpha_i \max_j(U_i - U_j, 0) - \beta_i \max_j(U_j - U_i, 0).$$

Every agent defecting is always an equilibrium. To see that cooperating is non-beneficial, notice that it results in lower utility U and inequity aversion -6β. Since both of these effects are negative, so is the total effect.

[Non-Uniqueness] For simplicity, take $x = 6$. An agent cooperates if and only if

$$6p + (1-p) - 6\beta(1-p) > 7p + 2(1-p) - 6\alpha p \Leftrightarrow$$

$$6\alpha p - 6\beta(1-p) > 1$$

Given a cooperation rate p, as a best response, a fraction $BR(p)$ of the population would wish to cooperate. We seek a fixed point, that is $BR(p) = p$, which will define an equilibrium. First, notice that if $p = 1/2$, the above equation becomes $3(\alpha - \beta) > 1$ which never holds because $\beta > \alpha$, so $BR(1/2) = 0$. On the other hand, if $p = 5/6$, then any agent who satisfies $5\alpha - \beta > 1$ will cooperate. So, take a distribution F with full support over Θ^{IA} such that $Pr(5\alpha - \beta > 1) = 0.99$. Then, $BR(5/6) = 0.99 > 5/6$. Because BR is continuous in p, there must exist some $1/2 < p' < 5/6$ such that $BR(p') = p'$, a non-trivial equilibrium. Recall that all agents defecting is also an equilibrium, so for such an F, the existence of at least two equilibria is now established. □

Proof of Proposition 3.6

Given the inequity aversion utility function, an agent cooperates if and only if

$$xp_x + (1-p_x) - 6\beta(1-p_x) > 7p_x + 2(1-p_x) - 6\alpha p_x$$

If $p_x > 0$, then it must be that some agent prefers $Coop$ to $Defect$. The RHS of the above inequality is decreasing in α and bounded above by β. So, the following is necessary:

$$\begin{aligned} xp_x + (1-p_x) - 6\beta(1-p_x) > 7p_x + 2(1-p_x) - 6\alpha p_x > 7p_x + 2(1-p_x) - 6\beta p_x \\ \Rightarrow \quad 6\beta(2p_x - 1) > p_x(6-x) + 1. \end{aligned}$$

The RHS of the last inequality above is positive, so cooperation can only be sustained if the LHS is also positive, that is $p_x > 1/2$. Therefore, any cooperation rate p_x where $1/2 > p_x > 0$ cannot be sustained by this model. Now, restricting attention to cooperation rates $p_x \geq 1/2$, we derive a tighter necessary condition. As the LHS is increasing in β and $\beta < 1$, it is necessary that

$$6(2p_x - 1) > 6\beta(2p_x - 1) > p_x(6-x) + 1$$

$$\Rightarrow \quad p_x > \frac{7}{6+x}$$

The argument is completed. □

Proof of Proposition 3.7

[Existence] Every player defecting is an equilibrium. This is because, for any player, switching to $Coop$ is harmful for both their material incentive U and the kindness payoff λK.

[Non-Uniqueness] We now provide an example for which there are infinitely many cooperative equilibria. Take the payoffs of Charness et al. (2016) with $x = 6$. Thus,

$$V_1(C,C) = 6 + \alpha \cdot \frac{3}{4} \qquad V_1(C,D) = 1 + \alpha \cdot \frac{-1}{4}$$

$$V_1(D,C) = 7 + \alpha \cdot \frac{-1}{4} \qquad V_1(D,D) = 2 + \alpha \cdot \frac{3}{4}$$

Therefore, if a p fraction of the population cooperates, a player's expected payoffs are:

$$\mathbb{E}\left[V_1(C,a_2)|p\right] = 6p + 1 \cdot (1-p) + \alpha \cdot \left(p - \frac{1}{4}\right)$$

$$\mathbb{E}\left[V_1(D,a_2)|p\right] = 7p + 2 \cdot (1-p) + \alpha \cdot \left(\frac{3}{4} - p\right)$$

Thus, an agent cooperates if and only if $-1 + \alpha(2p-1) \geq 0$. This implies that the cooperation rate p is sustainable in equilibrium if there is a $\hat{\alpha}$ such that $\hat{\alpha}(2p-1) = 1$ and $1 - F(\hat{\alpha}) = p$. Substituting in, it must be that $\hat{\alpha}(2(1 - F(\hat{\alpha})) - 1) = 1$. Hence, $F(\hat{\alpha}) = \frac{\hat{\alpha}-1}{2\hat{\alpha}}$. Now define F as such for $\hat{\alpha} \in [1,2]$ and freely elsewhere. Then, every $\hat{\alpha} \in [1,2]$ corresponds to an equilibrium with $p = 1 - F(\hat{\alpha}) = 1 - \frac{\hat{\alpha}-1}{2\alpha} = \frac{\hat{\alpha}+1}{2\hat{\alpha}}$. That is, every cooperation rate from 75% (when $\hat{\alpha} = 2$) to 100% (when $\hat{\alpha} = 1$) is an equilibrium with such an F. So we obtain a continuum of equilibria. As F was assigned freely outside of the range $[1,2]$, there could be other equilibria as well. Moreover, F could have been similarly assigned on a larger range so as to generate an even larger set of equilibrium. □

Proof of Proposition 3.8

Suppose that $1/2 \geq p_x > 0$. Then, it is the case that

$$\begin{aligned} p_x U_1(Defect, Coop) + (1-p_x)U_1(Defect, Defect) &> \\ p_x U_1(Coop, Coop) + (1-p_x)U_1(Coop, Defect) & \end{aligned}$$

and

$$\begin{aligned} \lambda p_x K(Defect, Coop) + \lambda(1-p_x)K(Defect, Defect) &\geq \\ \lambda p_x K(Coop, Coop) + \lambda(1-p_x)K(Coop, Defect) & \end{aligned}$$

Notice that plugging in the values of utility and K functions, and with terms matched up as shown, both these inequalities hold. Therefore, it is the case that:

$$\begin{aligned} & p_x V_1(Defect, Coop) \; + \; (1-p_x) \; V_1(Defect, Defect) > \\ & \quad p_x V_1(Coop, Coop) + \; (1-p_x) \; V_1 \; (Coop, Defect) \end{aligned}$$

So, every agent who chooses $Coop$ can profitably deviate to $Defect$, which is a contradiction to $p_x > 0$. □

Appendix II

Additional Analysis of Masatlioglu and Uler's (2013) Data

We focus on Masatlioglu and Uler's (2013) experimental design to the extent that is relevant for our study. We consider two types of questions. The first type is status quo-free choice problems where each subject is asked to make a choice between two alternatives, say x and y, i.e. $c(xy, \diamond)$ is observed for each subject. In the second type of questions, each subject is asked to imagine that he is given a status quo z that is *dominated by both* x *and* y; and he is asked to decide whether he wants to keep his status quo or change it for one of the other options x and y, i.e. $c(xyz, z)$ is observed for each subject (Figure 1). Each alternative is a bundle that is composed of some number of Guylian chocolate boxes and some amount of cash. At the end of the experiment, a question is randomly drawn for each subject and he receives the bundle that he has selected there.

We test the following property which is a modified version of the Status Quo Irrelevance axiom discussed in Section 2.

$$\text{If } c(xz, z) = x \text{ and } c(yz, z) = y, \text{ then } \; c(xy, \diamond) \; = \; c(xyz, z)$$

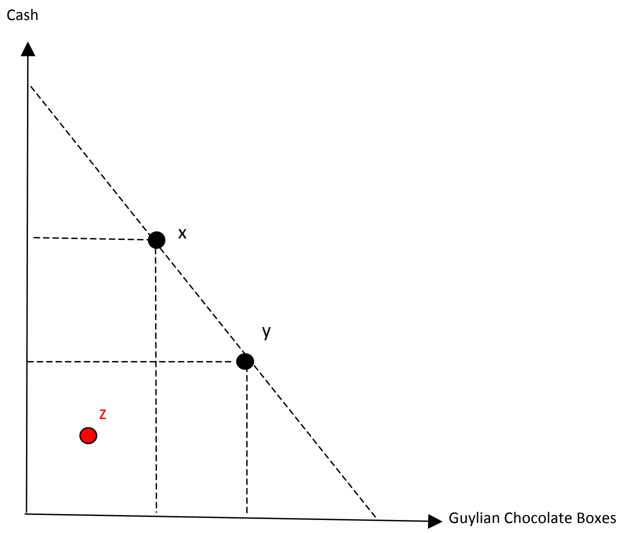

Figure 1: Bundles

According to the above property, z has no impact on the agent's choice between x and y when it is added into his choice set also with the feature that it is the status quo, because the agent is willing to leave z for both x and y. While this property must be satisfied by Masatlioglu and Ok (2005, 2014), it need not be the case in the switching cost model.

Since z is dominated by both x and y, a basic expectation is that z will not be chosen from choice problems (xz, z) and (yz, z). Thus, the if condition of the above property is expected to be naturally satisfied. Based on this expectation, Masatlioglu and Uler (2013) do not include questions of this type in their design and therefore the if condition cannot be directly verified in their data. However, they include questions of the type (xyz, z) and no subject chooses z there.

We test whether $c(xy, \diamond) = c(xyz, z)$ holds. When $c(xy, \diamond)$ is different from $c(xyz, z)$, we say that a choice reversal occurs at the status quo z. Otherwise, no reversal occurs at the status quo z. For each subject, there are three possible instances where a reversal may occur. We define a variable called REV to summarize the percentage of reversals each subject exhibits. The mean of REV is 31.66%. That is, 31.66% of the time, agents switch their choice from x to y or y to x upon the addition of a dominated status quo z into their choice problem. A Wilcoxon signed-rank test shows that this percentage is significantly different from zero (p=0.005). Therefore, the above property is violated significantly.[22]

[22] If these choice reversals are driven by a preference for randomness or are simply mistakes, then the frequency of reversals should be independent of the dominated status quo's location. Using McNemar test, we reject the hypothesis that reversals at two different status quos occur at the same rate ($p = 0.0082$).