Review of Economic Studies (2012) **79**, 375–387 doi: 10.1093/restud/rdr024

Advance access publication 18 August 2011

Eliciting Welfare Preferences from Behavioural Data Sets

ARIEL RUBINSTEIN
University of Tel Aviv Cafés, Tel Aviv University and New York University

and

YUVAL SALANT
Northwestern University

First version received September 2010; *final version accepted May* 2011 (*Eds.*)

An individual displays various preference orderings in different payoff-irrelevant circumstances. It is assumed that the variation in the observed preference orderings is the outcome of some cognitive process that distorts the underlying preferences of the individual. We introduce a framework for eliciting the individual's underlying preferences in such cases and then demonstrate it for two cognitive processes—satisficing and small assessment errors.

Key words: Welfare analysis, Preferences, Behavioural data set, Bounded rationality

JEL Codes: D01, D03, D60

1. INTRODUCTION

Our starting point is an individual whose behaviour in different circumstances is inconsistent with the maximization of a single preference ordering. Such inconsistency poses a challenge to welfare analysis as it is unclear from the individual's behaviour which preferences reflect his welfare. Our approach to addressing this challenge is based on two assumptions. The first is that the individual has underlying preferences that reflect his welfare. The second is that details that are irrelevant to the individual's interests systematically distort these preferences. Welfare analysis then requires the specification of the distortive mechanism and the identification of the individual's underlying preferences from his inconsistent behaviour.

Following is a concrete example.

Scenario I. An individual makes frequent choices from subsets of the set $\{a, b, c\}$ that are presented in alphabetical order. On Mondays, the individual chooses c whenever it is available and if it is not, then he chooses a. Thus, the individual's choice behaviour on Mondays is consistent with the maximization of the ordering $c \succ_1 a \succ_1 b$. On Fridays, the individual chooses a whenever it is available and if it is not, then he chooses c. Thus, the individual's choice behaviour on Fridays is consistent with the maximization of $a \succ_2 c \succ_2 b$. Does the individual prefer a over c or vice versa?

We suggest that in order to discuss the individual's preference between a and c, one needs to make assumptions as to why the individual behaves differently depending on the day, even though his underlying preferences are unchanged across days. One possibility is that the

individual is satisficing rather than maximizing his underlying preferences. In other words, he examines the alternatives in order and chooses the first alternative that is "good enough" in the sense that it exceeds some aspiration threshold that may vary across days. This satisficing conjecture is consistent with the data on choice behaviour for Mondays and Fridays. In fact, as shown in Section 2, the unique preference ordering that—together with the satisficing procedure—can "explain" the choice data is $\succ_1$ with only c exceeding the aspiration threshold on Mondays, and both a and c exceeding the threshold on Fridays, implying that the individual prefers c to a. $\diamondsuit$

In our framework, the *welfare* of an individual is reflected by an unobservable ordering (an asymmetric and transitive binary relation that relates every two alternatives) over a finite set X of feasible alternatives. What is observable is a behavioural data set that describes the individual's behaviour in different circumstances. A behavioural data set is a collection of orderings. In order to connect behaviour to the unobserved welfare preferences, a conjecture is made as to why the individual displays different preference orderings on different occasions. This conjecture is translated into a *distortion function* D that attaches to every ordering $\succ$ the set $D(\succ)$ of all orderings that may be displayed by an individual with the welfare ordering $\succ$.

For a given distortion function D, we say that an ordering $\succ$ is D-consistent with a behavioural data set Λ if every ordering in Λ is a distortion of $\succ$, *i.e.* if $\Lambda \subseteq D(\succ)$. If there exists a preference ordering that is D-consistent with Λ, we say that the behavioural data set Λ is D-consistent. For a given distortion function D, our goal is to identify conditions on a behavioural data set under which it is D-consistent. When a data set is D-consistent, we seek to characterize the set of preference orderings that are D-consistent with the data set.

For example, the classic model of rational choice, in which the decision-maker maximizes his underlying preferences, is captured by setting $D_R(\succ) = \{\succ\}$. In this model, a behavioural data set is D_R-consistent if and only if it is degenerate, *i.e.* it contains a single ordering. The next scenario is a less straightforward example of our approach.

Scenario II. An individual is asked to rank three political parties, L, C, and R. On one occasion, he expresses the Left view $L \succ_1 C \succ_1 R$ and on another the Center-Right view $C \succ_2 R \succ_2 L$. Thus, the behavioural data set is $\Lambda = \{\succ_1, \succ_2\}$. An observer conjectures that the inconsistency in the individual's reports is the outcome of errors made by the individual in assessing his preferences. Assessment errors are conjectured to be "small": they can change the relative ranking of two parties only when the parties are adjacent in the individual's underlying preference ordering. The distortion function that describes this small assessment errors process is

$$D_E(\succ) = \{\succ_f \ \mid \ x \succ y \succ z \text{ implies } x \succ_f z\}.$$

Because $D_E(C \succ L \succ R) = \{\succ_1, \succ_2, \succ\}$, the behavioural data set Λ is D_E-consistent. As shown in Section 3, the Center-Left view $C \succ L \succ R$ is the unique ordering that is D_E-consistent with Λ. $\diamondsuit$

Data on behaviour in our framework may be obtained in two ways. First, it may be obtained from occasional self-reports of an individual about his own preferences, such as in surveys or experiments. Second, it may be generated from observations on choice behaviour that is influenced by payoff-irrelevant details. Salant and Rubinstein (2008) model such choice behaviour using an extended choice function, which assigns a chosen alternative to every pair (A, f) where $A \subseteq X$ is a set of alternatives and f is a frame. A *frame* is a description of details that influence choice behaviour, though it is clear to an observer that they do not affect the individual's welfare. When an extended choice function satisfies standard consistency properties in a given frame, choice behaviour in each frame f can be explained by the maximization of a preference ordering $\succ_f$. In that case, a behavioural data set is the set of all preference orderings displayed in at least one frame.

In either interpretation of a behavioural data set, we postulate that any observed preference ordering is derived from the individual's welfare preferences according to a cognitive process described by a distortion function D. As is often the case, we may have several conjectures, *i.e.* distortion functions, as to why the individual expresses different preference orderings. In such cases, given a behavioural data set, we reject all conjectures with which the data set is not consistent and obtain a collection of candidate welfare orderings that are consistent with some unrefuted conjecture.

The dilemma of how to attach welfare preferences to a behavioural data set is related to the agenda of social choice theory: formulating "social welfare" preferences that aggregate the preference orderings of different individuals in a society. In a typical social choice exercise, desirable properties of an aggregation procedure are assumed and impossibility or possibility results are derived. Our setting is similar to that of the single-profile analysis in social choice theory (see *e.g.* Roberts (1980); Rubinstein (1984)) in the sense that our goal is to attach a welfare ordering to a single profile of orderings. In social choice theory, each of the orderings in a profile represents a different individual in the society, while in our framework, it represents the same individual in a different circumstance. The goal of social choice theory is to identify the society's welfare, while our goal was to uncover the individual's welfare.

We depart from standard social choice theory in two ways. First, we investigate potential cognitive deviations from an underlying welfare ordering rather than the aggregation of autonomic preference orderings of different individuals. In this sense, our approach can be thought of as an approach to social choice in which it is assumed that there exists a welfare ordering that reflects the common social interests, but individuals in the society make systematic mistakes in expressing these interests. Second, we study a framework in which the data are a set of orderings rather than a vector of orderings. Thus, we do not specify which frame results in a particular preference ordering nor do we account for whether the same preference ordering is expressed more than once. In the context of social choice, this is analogous to combining anonymity of individuals with invariance to the frequency of each preference ordering in the society.

Our approach to welfare analysis differs dramatically from the model-free Pareto approach to welfare advocated by Bernheim and Rangel (2007, 2009).[1] In the Pareto approach, an alternative a is Pareto-superior to an alternative b if a is ranked above b in all the observed preference orderings. The resulting Pareto relation[2] is typically a coarse binary relation that becomes even more so as the behavioural data set grows.[3] Our approach aims to make finer welfare judgements and to do so we make "testable" assumptions on the process that relates welfare preferences to behaviour. This makes it possible to infer the welfare ranking of two alternatives when one does not Pareto-dominate the other. In fact, applying our approach to some reasonable cognitive processes may even result in welfare judgements that are opposite to those of the Pareto relation, as the following scenario demonstrates.

Scenario III. When an individual is asked to rank the four alternatives a, b, c, and d, his responses in two different circumstances are $c \succ_1 b \succ_1 a \succ_1 d$ and $d \succ_2 b \succ_2 a \succ_2 c$. An observer conjectures that the individual may make mistakes in reporting his preferences and that if he

1. Apesteguia and Ballester (2010) also adopt a model-free approach to welfare. They define an index of how "far" choice observations are from maximizing a given preference ordering. Then, given a collection of choice observations, they interpret the preference ordering that minimizes that index as reflecting welfare.

2. Bernheim and Rangel (2007, 2009) call the Pareto relation the unambiguous choice relation.

3. See Manzini and Mariotti (2009) and Salant and Rubinstein (2008) for a critical discussion of the Pareto approach.

makes a "large" mistake and reverses the welfare ordering of two alternatives x and z, he also makes "smaller" mistakes and reverses the ordering of all the alternatives ranked between x and z according to his welfare preferences. In other words:

$$D_M(\succ) = \{\succ_f \mid \text{if } x \succ y \succ z \text{ and } z \succ_f x \text{ then } z \succ_f y \succ_f x\}.$$

In the behavioural data set $\Lambda = \{\succ_1, \succ_2\}$, the individual always ranks b above a, and thus it is tempting to conclude that he prefers b to a. But is it possible that he actually prefers a to b?

We now show that every ordering $\succ$ that is D_M-consistent with Λ ranks a above b. To see that, assume to the contrary that $b \succ a$ according to some ordering $\succ$ that is D_M-consistent with Λ. Note that: (1) if $d \succ b \succ a$ then $\succ_1 \notin D_M(\succ)$ and (2) if $b \succ a \succ d$ then $\succ_2 \notin D_M(\succ)$. Thus, $b \succ d \succ a$. Similarly $b \succ c \succ a$. We are left with two candidate welfare orderings: (1) $b \succ d \succ c \succ a$, but then $\succ_1 \notin D_M(\succ)$ and (2) $b \succ c \succ d \succ a$, but then $\succ_2 \notin D_M(\succ)$.

The two welfare orderings that are D_M-consistent with Λ are $c \succ a \succ b \succ d$ and $d \succ a \succ b \succ c$. Both rank a as welfare-superior to b, even though b Pareto-dominates a. $\diamondsuit$

Our approach to welfare analysis is in line with the recent choice theory literature on non-rational choice behaviour, *i.e.* behaviour that cannot be explained by the maximization of a single preference ordering.[4] In this literature, the data on behaviour is a choice function that specifies the individual's choices from each possible set of alternatives (rather than an extended choice function in which choices may depend on framing effects). When choice data are not consistent with the maximization of a preference ordering, this literature seeks to explain choice behaviour as an outcome of applying a decision-making procedure that takes as one of its parameters the individual's underlying preference ordering. For example, Manzini and Mariotti (2007, forthcoming), Cherepanov, Feddersen and Sandroni (2008), and Masatlioglu, Nakajima and Ozbay (2009) postulate a procedure of choice in which, given a set of alternatives, the decision-maker first identifies in some way a subset of alternatives to be considered. He then chooses from this consideration set according to his underlying preferences.

This choice theory literature characterizes conditions under which choice behaviour can be explained by the postulated procedure and then identifies parts of the individual's underlying preference ordering. Thus, as in our approach, one can interpret this literature as an attempt to elicit from choice observations the underlying preference ordering of the individual, and to do so, the literature makes assumptions about the individual's decision-making procedure. Green and Hojman (2007) adopt a related approach, in which the procedure of choice used by the decision-maker "aggregates" *several* conflicting considerations (each represented by an ordering). Given a choice function, Green and Hojman characterize the set of possible orderings that could have generated this function and apply the Pareto criterion to those orderings in order to make welfare judgements.

We proceed as follows. In Sections 2 and 3, we present two models that demonstrate our approach. The first model fits the interpretation of a behavioural data set as a description of choice behaviour, while the second fits the self-reporting interpretation. In each model, we begin by specifying the distortion function that describes the process by which the welfare preferences may be altered. We then identify conditions under which a behavioural data set is consistent with the cognitive process underlying the distortion function. When the data set is consistent, we extrapolate from it the set of candidate welfare orderings. Section 4 discusses possible modifications of our framework, and Section 5 concludes.

4. In fact, our approach does not rule out the possibility that choices do not necessarily reflect preferences even when behaviour is consistent with the maximization of a single preference ordering. Recent papers that point out to this possibility include Rubinstein (2006), Beshears *et al.* (2008), and Rubinstein and Salant (2008).

2. SATISFICING

The procedure of choice discussed in this section is Herbert Simon's (1955) Satisficing procedure. A satisficer has in mind some aspiration level, and he classifies an alternative as satisfactory or non-satisfactory depending on whether its value exceeds that level. In making choices, the alternatives are presented to the decision-maker in a predetermined order, such as alphabetical order. The decision-maker considers the alternatives in that order and chooses the first satisfactory alternative he encounters. If there are no satisfactory alternatives, he applies a "tie-breaking rule" to choose from among the non-satisfactory alternatives. We will examine two tie-breaking rules: a "perfect-recall" rule according to which the best non-satisfactory alternative is chosen and a "no-recall" rule in which the last alternative considered is chosen.

Satisficing behaviour may emerge when assessing the exact value of each alternative is difficult but determining whether an alternative is "good enough" is less so. This may be the case, for example, when considering candidates for a job. A short interview may enable a recruiter to provide a rough evaluation of the candidates and to choose a candidate who is good enough, if there is one. If there is no such candidate, a recruiter may settle for last interviewed candidate if the other candidates are no longer available (thus generating no-recall satisficing) or alternatively he may re-interview the candidates and choose the best one (thus generating perfect-recall satisficing). Satisficing may also emerge when there are search costs involved in considering an additional alternative that the decision-maker wishes to economize on. For example, when purchasing a product online, considering an additional alternative may be time consuming and a customer may therefore settle on an alternative that is good enough.[5]

Denote the order in which the alternatives are presented by O, where aOb means that alternative a is presented prior to alternative b. For a given aspiration level, a satisficer's choices are consistent with maximizing a unique ordering of the elements in X. This ordering positions all the satisfactory elements above all the non-satisfactory elements and ranks the satisfactory elements according to O. In the case of perfect-recall satisficing, the non-satisfactory elements are ranked according to the individual's welfare preferences, while in the case of no-recall satisficing, the non-satisfactory elements are ranked in opposite order to O. When aspiration levels vary according to, for example, the day of the week, a satisficer's choices will produce different rankings of the alternatives depending on the day and will thus result in a behavioural data set with more than one ordering.

Scenario I (continued). An individual makes choices from subsets of the set $\{a, b, c\}$, which are presented in alphabetical order. The individual's behaviour on Mondays is consistent with the maximization of the ordering $c \succ_1 a \succ_1 b$ and on Fridays with the maximization of $a \succ_2 c \succ_2 b$. The behavioural data set is thus $\{\succ_1, \succ_2\}$. We wish to determine whether choice behaviour can be explained by satisficing, and if so, to determine the welfare ordering of the alternatives under the satisficing hypothesis.

The individual's choice behaviour on Mondays cannot be explained by satisficing with no recall. If the individual were following no-recall satisficing, then his choice of c from the set $\{a, b, c\}$ would imply that a and b, which are considered prior to c, are not satisfactory. Thus, when choosing from the pair $\{a, b\}$ on Mondays, he should choose b but he actually chooses a.

The behavioural data set is consistent with perfect-recall satisficing. Consider, for example, an individual with the welfare preferences $c \succ a \succ b$ who finds only c to be satisfactory on Mondays but reduces his aspiration level on Fridays so that both a and c are satisfactory. Such

5. Tyson (2008) and Rubinstein and Salant (2006) discuss choice-theoretic aspects of satisficing, and Salant (2011) discusses procedural aspects of satisficing. Bendor (2003) surveys the use of satisficing in the political science literature to explain political phenomena.

an individual will choose c on Mondays whenever it is available and a otherwise. Thus, the individual's choices are consistent with the maximization of $\succ_1$ on Mondays and similarly with the maximization of $\succ_2$ on Fridays. In fact, the ordering $c \succ a \succ b$ is the unique welfare ordering consistent with the data set, as will be shown below. $\diamondsuit$

2.1. *Satisficing with perfect recall*

In satisficing with perfect recall, the decision-maker considers the alternatives in an order O and chooses the first satisfactory alternative he encounters. If there is no such alternative, he chooses the best available alternative according to his welfare preferences. The distortion function D_{PR} that describes the possible deviations of a Perfect-Recall satisficer from his welfare preferences assigns to every ordering $\succ$ a collection $D_{PR}(\succ)$ of orderings. An ordering $\succ_f$ is in $D_{PR}(\succ)$ if there exists a set $S \subseteq X$ such that:

(1a) every element in S is $\succ_f$-superior to every element in $X \setminus S$,
(1b) every element in S is $\succ$-superior to every element in $X \setminus S$,
(2) the $\succ_f$-ranking of the elements of S is according to O, and
(3) the $\succ_f$-ranking of the elements of $X \setminus S$ is according to $\succ$.

For example, if $X = \{a, b, c\}$ and the order of consideration is alphabetical, then $D_{PR}(c \succ a \succ b) = \{c \succ_1 a \succ_1 b, a \succ_2 c \succ_2 b, a \succ_3 b \succ_3 c\}$.

Given a behavioural data set Λ, we test the hypothesis that the decision-maker is a perfect-recall satisficer who uses the order O by examining whether there is an ordering $\succ$ such that $\Lambda \subseteq D_{PR}(\succ)$. If this is the case, we proceed to characterize the set of all orderings that could have generated the behavioural data set, *i.e.* the set $\{\succ \mid \Lambda \subseteq D_{PR}(\succ)\}$.

The key in the analysis is to define a binary relation $\succ_{PR}$ where $a \succ_{PR} b$ captures the intuitive inference that a must be welfare-superior to b given the behavioural data set and the satisficing hypothesis. For an ordering $\succ_f \in \Lambda$, let the Upper Tail of $\succ_f$, denoted $UT(\succ_f)$, be the largest set of elements at the top of $\succ_f$, which is ordered according to O. Let $LT(\succ_f) = X \setminus UT(\succ_f)$ be the Lower Tail of $\succ_f$. We define:

$$a \succ_{PR} b \text{ if there is } \succ_f \in \Lambda \text{ such that } a \succ_f b \text{ and } b \in LT(\succ_f).$$

The rationale behind this definition is that $b \in LT(\succ_f)$ implies that b is non-satisfactory in $\succ_f$ because the alternatives that are $\succ_f$-superior to b are not ordered according to O. Thus, any alternative that is $\succ_f$-superior to b is also welfare-superior to b.

The following proposition uses $\succ_{PR}$ to determine whether a given behavioural data set can be explained by perfect-recall satisficing. It also establishes that when a behavioural data set is D_{PR}-consistent, the collection of *all* D_{PR}-consistent preference orderings is the collection of all orderings that extend $\succ_{PR}$. In particular, $\succ_{PR}$ is the maximal binary relation nested in any candidate welfare ordering.[6]

Proposition 1. *For every behavioural data set* Λ*:*

(A) If the binary relation $\succ_{PR}$ *is cyclic then* Λ *is not* D_{PR}*-consistent. If* $\succ_{PR}$ *is asymmetric then* Λ *is* D_{PR}*-consistent.*
(B) A preference ordering is D_{PR}*-consistent with* Λ *if and only if it extends* $\succ_{PR}$*.*

6. A binary relation R is nested in a binary relation S if aRb implies that aSb. In this case, S extends R.

We first state two lemmas that simplify the proof of the proposition.

Lemma 1. *If $\succ_{PR}$ is asymmetric, then it is also acyclic.*

Proof. The asymmetry of $\succ_{PR}$ implies that for any two orderings $\succ_f$ and $\succ_g$ in Λ the following holds:

(i) Either $LT(\succ_f)$ is a subset of $LT(\succ_g)$ or vice versa. Otherwise, there are $a \in LT(\succ_f) \setminus LT(\succ_g)$ and $b \in LT(\succ_g) \setminus LT(\succ_f)$ such that $a \in UT(\succ_g)$ and $b \in UT(\succ_f)$ and thus $a \succ_{PR} b$ and $b \succ_{PR} a$.

(ii) The orderings $\succ_f$ and $\succ_g$ agree on the ranking of the elements in $LT(\succ_f) \cap LT(\succ_g)$. Otherwise, there are two elements $a, b \in LT(\succ_f) \cap LT(\succ_g)$ such that $a \succ_f b$ and $b \succ_g a$, implying that $a \succ_{PR} b$ and $b \succ_{PR} a$.

Let $\succ_h$ be the ordering in Λ with the largest lower tail. By (i) and (ii), this ordering is unique. The ordering $\succ_h$ nests $\succ_{PR}$. To see this, suppose that $a \succ_{PR} b$. Then, there exists $\succ_f \in \Lambda$ such that $a \succ_f b$ and $b \in LT(\succ_f)$. Because $LT(\succ_f) \subseteq LT(\succ_h)$, we obtain that $b \in LT(\succ_h)$. If $b \succ_h a$ then $a \in LT(\succ_h)$, and $b \succ_{PR} a$ in contradiction to the asymmetry of $\succ_{PR}$. Thus, $a \succ_h b$. We obtain that $\succ_h$ nests $\succ_{PR}$ and therefore $\succ_{PR}$ is acyclic. ‖

Lemma 2. *If an ordering $\succ$ is D_{PR}-consistent with Λ, then it extends $\succ_{PR}$.*

Proof. If $\succ$ is D_{PR}-consistent with Λ, then for every $\succ_f \in \Lambda$ there is a set $S(\succ_f) \subseteq X$ such that (1a) every element in $S(\succ_f)$ is $\succ_f$-superior to every element in $X \setminus S(\succ_f)$; (1b) every element in $S(\succ_f)$ is $\succ$-superior to every element in $X \setminus S(\succ_f)$; (2) the $\succ_f$-ranking of the elements of $S(\succ_f)$ is according to O; and (3) the $\succ_f$-ranking of the elements of $X \setminus S(\succ_f)$ is according to $\succ$.

Suppose that $a \succ_{PR} b$. Then, there is an ordering $\succ_f \in \Lambda$ such that $a \succ_f b$ and $b \in LT(\succ_f)$. By (2) above, all the elements in $S(\succ_f)$ belong to $UT(\succ_f)$ and thus $b \notin S(\succ_f)$. If $a \in S(\succ_f)$ then by (1b) above $a \succ b$, and if $a \notin S(\succ_f)$ then by (3) above $a \succ b$ since $a \succ_f b$. Thus, $a \succ b$, and we obtain that $\succ$ nests $\succ_{PR}$ as required. ‖

Proof of Proposition 1. To prove the first statement in part (A), suppose that $\succ_{PR}$ is cyclic. If Λ were D_{PR}-consistent, then there would be a preference ordering $\succ$ that is D_{PR}-consistent with Λ. By Lemma 2, the ordering $\succ$ would extend $\succ_{PR}$ in contradiction to $\succ_{PR}$ being cyclic.

To prove the second statement in part (A), suppose that $\succ_{PR}$ is asymmetric. By Lemma 1, $\succ_{PR}$ is also acyclic and thus can be extended to an ordering $\succ$. We now prove that $\succ$ is D_{PR}-consistent with Λ. Fix $\succ_f \in \Lambda$. To see that $\succ_f \in D_{PR}(\succ)$, we define $S(\succ_f) = UT(\succ_f)$ and verify that conditions (1a)–(3) in the definition of D_{PR} hold. Condition (1a) holds because every element in $S(\succ_f) = UT(\succ_f)$ is $\succ_f$-superior to every element in $X \setminus S(\succ_f) = LT(\succ_f)$. Condition (1b) holds because $a \succ_{PR} b$ for every $a \in S(\succ_f) = UT(\succ_f)$ and $b \in X \setminus S(\succ_f) = LT(\succ_f)$. By construction, the ordering $\succ$ nests $\succ_{PR}$ and thus $a \succ b$. Condition (2) holds because the elements in $UT(\succ_f)$ are ordered in $\succ_f$ according to O. Condition (3) holds because if $a, b \notin S(\succ_f)$ and $a \succ_f b$, then a and b are in $LT(\succ_f)$ and hence $a \succ_{PR} b$. This implies that $a \succ b$.

Part (B) follows from the proof of the second statement in part (A) and from Lemma 2. ‖

The following uniqueness result is an immediate corollary of Proposition 1.

Corollary 1. *Let Λ be a D_{PR}-consistent data set. There is a unique preference ordering that is D_{PR}-consistent with Λ if and only if there exists $\succ_f \in \Lambda$ such that the two $\succ_f$-maximal elements satisfy $a \succ_f b$ and bOa.*

Proof. Suppose Λ is D_{PR}-consistent. By part (B) of Proposition 1, there is a unique preference ordering that is D_{PR}-consistent with Λ if and only if there is a unique ordering that extends $\succ_{PR}$, that is, if and only if $\succ_{PR}$ is connected. By the proof of Lemma 1, the relation $\succ_{PR}$ is connected if and only if the lower tail of some ordering $\succ_f \in \Lambda$ contains $|X|-1$ elements, which is equivalent to the stated condition. ‖

2.2. *Satisficing with no recall*

In satisficing with no recall, the decision-maker chooses the first satisfactory alternative he encounters and if there is no such alternative, the last available alternative. The distortion function D_{NR} that describes the possible deviations of a No-Recall satisficer from his welfare preferences assigns to every ordering $\succ$ a collection $D_{NR}(\succ)$ of orderings. An ordering $\succ_f$ is in $D_{NR}(\succ)$ if there exists a set $S \subseteq X$ such that:

(1a) every element in S is $\succ_f$-superior to every element in $X \setminus S$,
(1b) every element in S is $\succ$-superior to every element in $X \setminus S$,
(2) the $\succ_f$-ranking of the elements of S is according to O, and
(3) the $\succ_f$-ranking of the elements of $X \setminus S$ is according to the reverse of O.

As before, the key to the analysis is to define a relation $\succ_{NR}$ that captures the welfare inferences that can be made from the data. Let Z be the O-minimal element in X. Given a behavioural data set Λ, we define:

$$a \succ_{NR} b \text{ if there is } \succ_f \in \Lambda \text{ such that } a \succ_f Z \text{ and } Z \succ_f b.$$

The rationale for this definition is that $a \succ_f Z$ implies that a is satisfactory, whereas $Z \succ_f b$ implies that b is not. Note that the relation $\succ_{NR}$ is silent as to how Z relates to other elements.

In order to state the next result, we need the following definition. Given an ordering P of the elements of X, we say that the element Z is the O-single trough of P if all the elements that are P-superior to Z are ordered in P as they are in O and all the elements that are P-inferior to Z are ordered in P in the reverse order to O. That is, for every two elements $a, b \in X$, $aPbPZ$ implies $aObOZ$ and $ZPaPb$ implies $bOaOZ$.

The following proposition shows that if $\succ_{NR}$ is cyclic or if the element Z is not the O-single trough of some ordering in the behavioural data set, then choice behaviour cannot be explained by satisficing with no recall. It also establishes that when $\succ_{NR}$ is acyclic and Z is the O-single trough of every ordering in the behavioural data set, then choice behaviour can be explained by no-recall satisficing. Moreover, the collection of *all* D_{NR}-consistent preference orderings is the collection of all orderings that extend $\succ_{NR}$.

Proposition 2. *For every behavioural data set Λ:*

(A) The data set Λ is D_{NR}-consistent if and only if (i) the relation $\succ_{NR}$ is acyclic and (ii) the element Z is the O-single trough of every ordering in Λ.
(B) If the data set Λ is D_{NR}-consistent, then a preference ordering is D_{NR}-consistent with Λ if and only if it extends the relation $\succ_{NR}$.

Proof. *(A)* Suppose a behavioural data set Λ is D_{NR}-consistent and let $\succ$ be an ordering that is D_{NR}-consistent with Λ. Then, for every $\succ_f \in D_{NR}(\succ)$, there exists a set $S(\succ_f) \subseteq X$ such that (1a) every element in $S(\succ_f)$ is $\succ_f$-superior to every element in $X \setminus S(\succ_f)$, (1b) every element in $S(\succ_f)$ is $\succ$-superior to every element in $X \setminus S(\succ_f)$, (2) the $\succ_f$-ranking of the elements of $S(\succ_f)$ is according to O, and (3) the $\succ_f$-ranking of the elements of $X \setminus S(\succ_f)$ is according to the reverse of O.

To prove (i), it is sufficient to show that $\succ$ nests $\succ_{NR}$. Suppose that $a \succ_{NR} b$. Then, there is an ordering $\succ_f \in \Lambda$ such that $a \succ_f Z \succ_f b$. If $Z \in S(\succ_f)$ then by (1a) above $a \in S(\succ_f)$ and by (2) above $b \in X \setminus S(\succ_f)$. Thus, by (1b) $a \succ b$. If $Z \notin S(\succ_f)$ then by (3) above $a \in S(\succ_f)$, by (1a) above $b \in X \setminus S(\succ_f)$ and thus by (1b) $a \succ b$.

To prove (ii), consider an ordering $\succ_f \in \Lambda$. If $Z \in S(\succ_f)$ then by (2) above all elements that are $\succ_f$-superior to Z are ordered according to O, and by (3) above all elements that are $\succ_f$-inferior to Z are in $X \setminus S(\succ_f)$ and their ordering is according to the reverse of O. Thus, Z is the O-single trough of $\succ_f$. A similar argument holds if $Z \in X \setminus S(\succ_f)$.

Suppose now that (i) and (ii) hold. Let $\succ$ be an ordering that extends $\succ_{NR}$. For every $\succ_f \in \Lambda$, define the set $S(\succ_f) = UT(\succ_f) \setminus \{Z\}$ and add Z to $S(\succ_f)$ if $Z \succ x$ for all $x \in LT(\succ_f)$. Then:

(1a) holds because $S(\succ_f)$ is either the upper tail of $\succ_f$ or the upper tail of $\succ_f$ excluding Z, which is the $\succ_f$-minimal element in the upper tail. In either case, every alternative in $S(\succ_f)$ is $\succ_f$-superior to every element in $X \setminus S(\succ_f)$;

(1b) holds because $\succ$ extends $\succ_{NR}$, and $a \succ_{NR} b$ if $a \in S(\succ_f) \setminus \{Z\}$ and $b \in X \setminus S(\succ_f) \setminus \{Z\}$. In addition, if $Z \in S(\succ_f)$ then Z is $\succ$-superior to every element in $X \setminus S(\succ_f)$ by construction, and if $Z \notin S(\succ_f)$ then there exists $b \notin S(\succ_f)$ such that $b \succ Z$ and b is $\succ$-inferior to all elements in $S(\succ_f)$ implying the same for Z;

(2) and (3) hold because the ranking of elements in $UT(\succ_f)$ is according to O and the ranking of elements in $LT(\succ_f)$ is according to the reverse of O.

(B) Follows from the the proof of part (A). ‖

Note that in satisficing with no recall, the welfare ranking of Z is never identified from choice data. In fact, by Proposition 2, if a preference ordering $\succ$ is D_{NR}-consistent with a behavioural data set, then so is any ordering obtained from $\succ$ by any change in the position of Z.

Note also that if a behavioural data set is consistent with both versions of satisficing, then the set of welfare orderings that are consistent with perfect-recall satisficing is a subset of the set of welfare orderings that are consistent with no-recall satisficing. Intuitively, this is because the tie-breaking rule in satisficing with perfect recall imposes more restrictions on the link between welfare preferences and choice behaviour than the tie-breaking rule in the no-recall case.[7]

3. SMALL ASSESSMENT ERRORS

The discussion in this section fits the interpretation of a behavioural data set as a collection of orderings self-reported by an individual in various circumstances that an observer believes do not influence the individual's welfare.

Consider an individual who views the alternatives of X as evenly spread out along the utility spectrum, such that the distance between every pair of adjacent alternatives is similar. When reporting his preferences, the individual may overestimate or underestimate the value of any given alternative. This may be due to, for example, the complexity of the alternatives or the difficulty in detecting minor details. Assessment errors are "small" in the sense that the size of an error is less than the utility distance between two adjacent alternatives. That is, assessment errors change the ordering of two alternatives only when the alternatives are adjacent in the individual's underlying preference ordering, the higher one is underestimated and the lower one is overestimated.

7. Formally, assume that a data set Λ is D_{PR}-consistent and D_{NR}-consistent. By Propositions 1 and 2, it is sufficient to show that the relation $\succ_{PR}$ defined in perfect-recall satisficing nests the relation $\succ_{NR}$ defined in no-recall satisficing. Assume $a \succ_{NR} b$. Then, there exists an ordering $\succ_f \in \Lambda$ such that $a \succ_f Z \succ_f b$. This implies that $b \in LT(\succ_f)$ and since $a \succ_f b$ we obtain that $a \succ_{PR} b$.

Formally, let D_E be the distortion function that attaches to every preference ordering $\succ$ all the orderings that are obtained from $\succ$ by disjoint switches of $\succ$-adjacent alternatives. In other words, $D_E(\succ) = \{\succ_f \mid a \succ b \succ c \text{ implies } a \succ_f c\}$.

To examine whether small assessment errors can generate a given behavioural data set Λ, we define $a \succ_E b$ if there exists $\succ_f$ in Λ and an element x such that $a \succ_f x \succ_f b$. The following proposition establishes that when the behavioural data set is D_E-consistent, any extension of $\succ_E$ to an ordering is D_E-consistent with the data set.

Proposition 3. *For every behavioural data set Λ:*

(A) If the relation $\succ_E$ is cyclic then Λ is not D_E-consistent. If $\succ_E$ is 3-acyclic then Λ is D_E-consistent.[8]

(B) An ordering is D_E-consistent with Λ if and only if it extends $\succ_E$.

Proof. We prove part (A). Part (B) immediately follows.

To prove the first statement in part (A), we show that if $\succ$ is an ordering that is D_E-consistent with Λ, then $\succ$ nests $\succ_E$ and therefore $\succ_E$ is acyclic. Suppose that $a \succ_E b$. Then, there exists $\succ_f \in \Lambda$ and an element x such that $a \succ_f x \succ_f b$. Assume to the contrary that $b \succ a$. If $b \succ x \succ a$, we would not have $a \succ_f b$. If $x \succ b \succ a$, we would not have $a \succ_f x$. If $b \succ a \succ x$, we would not have $x \succ_f b$. Thus, since $\succ$ relates every two alternatives, we obtain that $a \succ b$.

To prove the second statement in part (A), we first show that if $\succ_E$ is 3-acyclic, it is also acyclic. Suppose $\succ_E$ has a cycle and consider the *shortest* one $x_1 \succ_E x_2 \succ_E \ldots \succ_E x_K \succ_E x_1$. Since $\succ_E$ is 3-acyclic, we have that $K > 3$. Because $x_1 \succ_E x_2$, there is $\succ_f$ in Λ such that $x_1 \succ_f x \succ_f x_2$. There exists $k \in \{3, 4\}$ such that x is not equal to x_k. If $x_k \succ_f x$, then by definition $x_k \succ_E x_2$, and if $x \succ_f x_k$ then $x_1 \succ_E x_k$. In either case, we obtain a shorter cycle.

Because the relation $\succ_E$ is acyclic, it can be extended to an ordering $\succ$. Assume to the contrary that $\succ$ is not D_E-consistent with Λ. Then, there are three elements a, x and b such that $a \succ x \succ b$ and $b \succ_f a$ for some $\succ_f \in \Lambda$. This cannot occur since (i) if $x \succ_f b \succ_f a$ then $x \succ_E a$ contradicting $a \succ x$, (ii) if $b \succ_f x \succ_f a$ then $b \succ_E a$ contradicting $a \succ b$, and (iii) if $b \succ_f a \succ_f x$ then $b \succ_E x$ contradicting $x \succ b$. ‖

There may be more than one preference ordering that is D_E-consistent with a given behavioural data set. For example, if the data set contains only one ordering, then any ordering of the alternatives obtained from that ordering by disjoint switches of adjacent elements is D_E-consistent with the data set. Proposition 4 identifies a necessary and sufficient condition for the existence of a unique preference ordering that is D_E-consistent with the data set.

Proposition 4. *Assume that a behavioural data set Λ is D_E-consistent. There exists a unique preference ordering that is D_E-consistent with Λ if and only if for every pair of elements a and b at least one of the following holds:*

(i) There is an ordering $\succ_f$ in Λ and an alternative x such that x is ranked between a and b in $\succ_f$,

(ii) There are two orderings in Λ and an alternative x such that according to one of the orderings x is ranked above both a and b and according to the other x is ranked below both of them.

8. A binary relation S is 3-acyclic if it does not contain cycles of three or fewer elements.

Proof. The "if" part: Since $\succ$ nests $\succ_E$, it is sufficient to show that $\succ_E$ is connected. Fix two alternatives a and b. If (i) holds, then $\succ_E$ relates a and b. If (ii) holds, then there are two orderings in Λ, $\succ_f$ and $\succ_g$, and an element x such that x is $\succ_f$-superior to both a and b and $\succ_g$-inferior to both a and b. Suppose (without loss of generality) that $a \succ_f b$. Then $a \succ_g b$ since otherwise $x \succ_f a \succ_f b$ and $b \succ_g a \succ_g x$ would imply that $x \succ_E b$ and $b \succ_E x$. Thus, we have that $x \succ_f a \succ_f b$ and $a \succ_g b \succ_g x$. Therefore, $x \succ_E b$ and $a \succ_E x$ and the relation $\succ_E$ connects a and b.

The "only if" part: Consider two alternatives a and b such that both (i) and (ii) do not hold. Let U (D) denote the set of elements that are above (below) both a and b in all the orderings in Λ. Since (i) and (ii) do not hold, the sets U and D are disjoint and contain all the elements of X other than a and b. We now show that $\succ_E$ does not connect a and b, and hence by Proposition 3 both orderings of these two alternatives are possible. Assume to the contrary that $a \succ_E x_1 \succ_E x_2 \succ_E \ldots \succ_E x_k \succ_E b$. Then, $x_k \in U$ because $x_k \succ_E b$ implies that there is an ordering $\succ_f$ and an element y such that $x_k \succ_f y \succ_f b$. By applying similar arguments, we obtain that $x_1 \in U$. However, by $a \succ_E x_1$, we have that x_1 is ranked two or more places below a in some ordering in Λ and thus $x_1 \in D$ in contradiction to U and D being disjoint. ‖

4. POSSIBLE MODIFICATIONS OF THE FRAMEWORK

The data on behaviour in our framework are a set of orderings. This fits situations in which an observer does not have information on the actual frames that triggered the individual to behave inconsistently. This also fits situations in which an observer does not have a theory specifying exactly how each frame distorts the underlying preferences. An alternative framework would be one in which the observer has such a theory.

To formally describe this modified framework, we define an *extended behavioural data set* to be a set $\Lambda = \{(\succ_f, f)\}$ where $\succ_f$ is an ordering and f is a frame. A frame f is additional information regarding the circumstances in which the individual displays the preference ordering $\succ_f$. The effect of each frame on the individual's welfare preferences is summarized by a distortion function D that attaches a set of orderings to every pair $(\succ, f)$ where $\succ$ is an ordering and f is a frame. An ordering $\succ$ is D-consistent with Λ if $\succ_f \in D(\succ, f)$ for every $(\succ_f, f) \in \Lambda$.

The same ordering can appear multiple times in an extended behavioural data set and be associated with different frames. The operation of attaching a welfare ordering to an extended behavioural data set is analogous to the single-profile social choice exercise without assuming anonymity or invariance to the frequency of each preference ordering in the society. The assumptions behind the distortion function play a role that is analogous to that of a set of axioms in the social choice theory.

Scenario IV. An external mechanism highlights some of the alternatives in X. For example, a website presents some of the alternatives in a special colour, and a grocery store positions some products near the cashier. An individual is influenced by this highlighting: he attaches a non-negative "bonus" to every highlighted alternative and improves its ranking with respect to non-highlighted alternatives. The ordering of the non-highlighted alternatives remains according to the underlying preference ordering, while the ordering of the highlighted alternatives may change. Formally, a frame f is a subset of the set of feasible alternatives X and

$$D_H(\succ, f) = \{\succ_f | \text{ if } [b \succ a \text{ and } a \succ_f b] \text{ then } a \in f\}.$$

When observing *both* a preference ordering $\succ_f$ and a set of highlighted elements f, one can infer the welfare ordering between a non-highlighted alternative and any alternative that is $\succ_f$-inferior to it. Thus, given an extended behavioural data set Λ, we define $a \succ_H b$ if there exists a

pair $(\succ_f, f)$ such that $a \succ_f b$ and $a \notin f$. It is straightforward to show that the set of orderings that are D_H-consistent with Λ is the set of all orderings that extend $\succ_H$.[9]$\diamondsuit$

Another possible modification of our framework relates to the ordinality of welfare preferences. It is sometimes natural to refer also to the intensity of the preferences when describing the cognitive process that distorts them. For example,

Scenario V. An individual is influenced by advertising. He may prefer product a to product b, but the number of times he views advertisements for each product may influence his choice between them. In order to describe the magnitude of the advertising bias, we need to introduce a notion of cardinal utility.

Formally, an advertising frame is a function $i : X \to N$ that assigns to every alternative $x \in X$ the number of advertisements $i(x)$ for that alternative. An individual is influenced by advertising: he has in mind a welfare utility function u that assigns positive values to the different alternatives, yet he maximizes $i(x)u(x)$ in frame i instead of his utility $u(x)$.

Given an extended behavioural data set $\Lambda = \{(\succ_i, i)\}$, an ordering $\succ$ is consistent with Λ if there is a utility representation u of $\succ$ such that for every pair $(\succ_i, i)$ the function $i(x)u(x)$ represents $\succ_i$. The existence of such a function u is equivalent to the existence of a solution to a system of inequalities in $|X|$ unknowns $\{u(x)\}_{x\in X}$, where each inequality is of the form $i(x)u(x) > i(y)u(y)$ for $x \succ_i y$.$\diamondsuit$

5. CONCLUSION

This paper refers to situations in which the same individual displays different preference orderings in various circumstances that differ in payoff-irrelevant parameters. An observer conjectures that this is the result of systematic deviations from an underlying preference ordering, which respects the individual's welfare, and wishes to elicit that ordering from data on the individual's behaviour. In the previous sections, we illustrated the elicitation process in several scenarios.

The distortion function in our framework is deterministic in the sense it does not specify how likely an individual with a particular welfare ordering is to express various orderings. A related framework would be one in which the distortion function assigns to every welfare ordering a probability measure over orderings. A richer set of questions can then be analysed. For example, given a cognitive process, it is possible to determine which welfare ordering is "most likely" to have generated the behavioural data set, and given several candidate cognitive processes, it is possible to determine the "fit" of each process and select the process with the best fit. A branch of social choice theory, which traces back to Condorcet (1785), follows a similar approach. This approach assumes that all orderings expressed by individuals in a society are stochastic distortions of a true social welfare ordering and aims to characterize the most likely social welfare ordering given certain assumptions on the nature of the stochastic distortion. Examples of this approach include Nitzan and Paroush (1985), Young (1988) and more recently Baldiga and Green (2009).

Our approach to welfare analysis in the presence of behavioural biases highlights the dilemma of an observer who wishes to attach welfare preferences to an individual who behaves inconsistently. The observer may solve this dilemma differently in different contexts since he may have different conjectures about the cause for the inconsistent behaviour. The essence of our approach

9. We first prove that $\succ_H$ is nested in any ordering $\succ$ that is D_H-consistent with Λ. Suppose that $a \succ_H b$. Then, there is a frame f such that $a \succ_f b$ and $a \notin f$. To see that $a \succ b$, note that if $b \succ a$, then $a \succ_f b$ and $b \succ a$ would imply that $a \in f$. We now prove that if an ordering $\succ$ extends $\succ_H$, then it is D_H-consistent with Λ. We need to verify that if $b \succ a$ and $a \succ_f b$, then $a \in f$. Otherwise, $a \notin f$ and $a \succ_f b$ imply that $a \succ_H b$ contradicting the assumption that $\succ$ extends $\succ_H$.

is that making meaningful welfare inferences requires making assumptions on the mapping from preferences to behaviour. We demonstrated throughout the paper that developing a welfare concept based on such assumptions is analytically tractable.

Acknowledgment. tnqdeleteWe thank Ayala Arad, Doug Bernheim, Eddie Dekel, Jerry Green, Daniel Hojman, Tim Feddersen, Drew Fudenberg, Yusufcan Masatlioglu, Meg Meyer, Ron Siegel, Ran Spiegler, Rakesh Vohra, two anonymous referees, and seminar participants at Boston University, Hebrew University, Harvard University, New York University, Northwestern University, University of British Columbia, the Summer School on Welfare Economics and Philosophy at San-Sebastian (Spain, July 2009), and the European Summer Symposium in Economic Theory at Gerzensee (Switzerland, July 2010) for their comments. Rubinstein acknowledges support from the Israeli Science Foundation (259/08) and from the Foerder Institute for Economic Research.

REFERENCES

APESTEGUIA, J. and BALLESTER, M. A. (2010), "A Measurement of Rationality and Welfare" (Economics Working Paper No. 1220, Department of Economics and Business, Universitat Pompeu Fabra).
BALDIGA, K. A. and GREEN, J. R. (2009), "Choice-Based Measures of Conflict in Preferences". (Mimeo, Harvard University).
BENDOR, J. (2003), "Herbert A. Simon: Political Scientist", *Annual Review of Political Science*, **6**, 433–471.
BERNHEIM, B. D. and RANGEL, A. (2007), "Toward Choice-Theoretic Foundations for Behavioral Welfare Economics", *American Economic Review Papers and Proceedings*, **97**, 464–470.
BERNHEIM, B. D. and RANGEL, A. (2009), "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics", *Quarterly Journal of Economics*, **124**, 51–104.
BESHEARS, J., CHOI, J. J., LAIBSON, D. and MADRIAN, B. C. (2008), "How are Preferences Revealed?", *Journal of Public Economics*, **92**, 1787–1794.
CHEREPANOV, V., FEDDERSEN, T. and SANDRONI, A. (2008), "Rationalization" (Mimeo, Northwestern University).
CONDORCET, M. (1785), "Essai surl'application de l'analyse à la probabilité des décisions rendues à la probabilité des voix". (Paris: De l'imprimerie royale).
GREEN, J. R. and HOJMAN, D. A. (2007), "Choice, Rationality and Welfare Measurement" (Harvard Institute of Economic Research Discussion Paper No. 2144 and KSG Working Paper No. RWP07-054).
MANZINI, P. and MARIOTTI, M. (2007), "Sequentially Rationalizable Choice", *American Economic Review*, **97**, 1824–1839.
MANZINI, P. and MARIOTTI, M. (2009), "Choice Based Welfare Economics for Boundedly Rational Agents". (Mimeo, University of St Andrews).
MANZINI, P. and MARIOTTI, M. (forthcoming), "Categorize Then Choose: Boundedly Rational Choice and Welfare", *Journal of the European Economic Association.*
MASATLIOGLU, Y., NAKAJIMA, D. and OZBAY, E. Y. (2009), "Revealed Attention". (Mimeo, University of Michigan).
NITZAN, S. and PAROUSH, J. (1985), Collective Decision Making: An Economic Outlook (London and New York: Cambridge University Press).
ROBERTS, K. W. S. (1980), "Social Choice Theory: The Single-Profile and Multi-Profile Approaches", *Review of Economic Studies*, **47**, 441–450.
RUBINSTEIN, A. (1984), "The Single Profile Analogies to Multi Profile Theorems: Mathematical Logic's Approach", *International Economic Review*, **25**, 719–730.
RUBINSTEIN, A. (2006), "Comments on Behavioral Economics", in Blundell, R., Newey, W. K. and Persson, T. (eds) *Advances in Economic Theory (2005 World Congress of the Econometric Society)*, Vol. 2. (Cambridge: Cambridge University Press) 246–254.
RUBINSTEIN, A. and SALANT, Y. (2006), "A Model of Choice from Lists", *Theoretical Economics*, **1**, 3–17.
RUBINSTEIN, A. and SALANT, Y. (2008), "Some Thoughts on the Principle of Revealed Preference", in Caplin, A and Schotter, A. (eds) *Handbooks of Economic Methodologies* (New York: Oxford University Press) 115–124.
SALANT, Y. (2011), "Procedural Analysis of Choice Rules with Applications to Bounded Rationality", *American Economic Review*, **101**, 724–748.
SALANT, Y. and RUBINSTEIN, A. (2008), "(A, f): Choice with Frames", *Review of Economic Studies*, **75**, 1287–1296.
SIMON, H. A. (1955), "A Behavioral Model of Rational Choice", *Quarterly Journal of Economics*, **69**, 99–118.
TYSON, C. (2008), "Cognitive Constraints, Contraction Consistency, and the Satisficing Criterion", *Journal of Economic Theory*, **138**, 51–70.
YOUNG, H. P. (1988), "Condorcet's Theory of Voting", *American Political Science Review*, **82**, 1231–1244.