Eliciting Welfare Preferences from Behavioral Datasets

Ariel Rubinstein
University of Tel Aviv Cafés,
Tel Aviv university,
New York University

and

Yuval Salant
Northwestern University

February 9, 2010

Abstract. A behavioral dataset contains various preference relations displayed by a single individual in different payoff-irrelevant circumstances. We introduce a framework for eliciting the individual's welfare preferences in such cases. In this framework, the different preference relations displayed by the individual are the outcome of a cognitive process that distorts an unobserved preference relation reflecting the individual's welfare. We demonstrate the operation of eliciting the underlying welfare preferences from behavioral datasets for several cognitive processes.

We thank Ayala Arad, Doug Bernheim, Daniel Hojman, Tim Feddersen, Drew Fudenberg, Yusufcan Masatlioglu, Rani Spiegler, Rakesh Vohra, Asher Wolinsky and seminar participants at Hebrew University, Northwestern University, UBC and the Summer School on Welfare Economics and Philosophy at San-Sebastian (Spain, July 2009) for their comments.

1. Introduction

Welfare analysis requires the specification of a welfare criterion. In economic models where an individual has clear preferences that guide his choices, these preferences are usually assumed to reflect his welfare. However, in many real-life situations, an individual displays different preferences in different frames, where a *frame* includes details that appear to be irrelevant to the individual's own interests (see Salant and Rubinstein (2008)). Frames may be external, a consequence of the environment in which the individual acts, or internal, a consequence of the individual's state of mind. Frames may affect the cognitive process by which an individual evaluates alternatives, and as a result the individual may display different preferences in different frames.[1] In such cases, there is no straightforward method to determine the individual's welfare preferences.

Consider, for example, an individual who is busier on some days of the week than on others. In making choices, the individual examines the alternatives one after the other and chooses the first alternative that he finds satisfactory. On days in which he is busier, the individual is more easily satisfied and hence behaves differently than on days in which he is less so. What can we infer about the individual's welfare ranking of the different alternatives from his "inconsistent" choice behavior? Similarly, when repeatedly surveying a shopper to learn about his tastes, we may find that they are different in each survey. We might conjecture that his underlying preferences are in fact stable across surveys but that consumer trends are affecting his responses. Can we elicit his preferences in the absence of the trend distortion? A similar question may arise in the context of organizations. We may hear different members of an organization express somewhat different goals for the organization. We might conjecture that they all share the same underlying organizational goals, and that the differences in their views result from small misunderstandings or miscommunication. Can we elicit the actual underlying goals of the organization?

The goal of this paper is to present a framework for eliciting welfare preferences when an individual is affected by framing, and demonstrate the operation of eliciting these preferences for several cognitive processes. In our framework, the *welfare* of an individual is reflected by a unique preference relation[2] over a set X of N feasible alternatives. While the welfare relation is not observable, a set of preference relations, called a *behavioral dataset*, is. The behavioral dataset contains preference relations that are postulated to be systematic deviations from the welfare relation. All possible deviations are described by a consistency function C that attaches to every welfare relation $\succ$ the set of all preference relations $C(\succ)$ that may be displayed by the individual in some frame. The consistency function C specifies the possible outputs of the decision maker's cognitive process in different frames.

Given a behavioral dataset Λ and a consistency function C, we say that a preference relation $\succ$ is C-consistent with Λ if every relation in Λ is a distortion of $\succ$, i.e. it appears in $C(\succ)$. If there exists a preference relation that is C-consistent with Λ, we say that the behavioral dataset Λ is C-consistent. For a given consistency function C, we identify conditions on a behavioral dataset under which it is C-consistent. When a dataset is C-consistent, we characterize the set of preference

[1] For a choice-theoretic analysis of frame-dependent choice, see Salant and Rubinstein (2008). Bernheim and Rangel (2007, 2009) independently discuss a similar framework called choice with ancillary conditions.

[2] A preference relation is a complete, asymmetric and transitive binary relation.

relations that are C-consistent with the dataset. For example, in the standard approach to welfare, a decision maker has clear preferences that he maximizes in making choices and thus $C(\succ) = \{\succ\}$, i.e. the decision maker never distorts his underlying preferences. In this case, if an individual displays two distinct preference relations, then he is not C-consistent.

An observation in our framework is a full preference relation. This fits situations like surveys in which a decision maker expresses a complete ranking of the alternatives in each survey. This also fits situations in which a researcher observes a decision maker choosing inconsistently from choice problems, and notices that choice behavior can be explained by the maximization of several preference relations, each applied in a different frame. In both cases, we postulate that any of the observed preference relations relates to the welfare preferences as described by the consistency function C but lack a theory explaining how a specific frame distorts the welfare preferences.

Our approach to welfare analysis in the presence of framing effects differs from the Pareto approach advocated by Bernheim and Rangel (2007, 2009).[3] In the Pareto approach, an alternative a is Pareto-superior to an alternative b if all the observed preference relations rank a above b. The resulting Pareto relation is typically a coarse binary relation that becomes even coarser the larger the behavioral dataset.[4]

We assert that making explicit assumptions on the process that relates welfare preferences to behavior is necessary in order to elicit an individual's welfare preferences. By making such assumptions, one may sometimes be able to infer the welfare ranking of two alternatives in cases where one does not Pareto-dominate the other. In fact, applying our approach to some reasonable cognitive processes may even result in welfare rankings that are opposite to those of the Pareto relation. The following example demonstrates such a case.

Large mistakes result in small mistakes. Each observed preference relation satisfies the following property: If the decision maker makes a "large" mistake by reversing the welfare ranking of two elements x and z, he also makes "smaller" mistakes by reversing the welfare ranking of x and y and of y and z for every element y that lies between x and z according to his welfare preferences. Thus $C(\succ) = \{\succ_f \mid \text{if } x \succ y \succ z \text{ and } z \succ_f x \text{ then } z \succ_f y \succ_f x\}$.

Let $X = \{a, b, c, d\}$ and assume that the behavioral dataset contains the following two orderings:

$(\succ_f)$ $c \succ_f b \succ_f a \succ_f d$ and

$(\succ_g)$ $d \succ_g b \succ_g a \succ_g c$.

Although b Pareto-dominates a, every relation $\succ$ that is C-consistent with $\Lambda = \{\succ_f, \succ_g\}$ ranks a above b. To see this, assume to the contrary that $b \succ a$ according to some relation $\succ$ that is C-consistent with Λ. Then d is ranked between b and a in $\succ$ (i.e. $b \succ d \succ a$) since: (i) if d is welfare-superior to b then $\succ_f \notin C(\succ)$ and (ii) if d is welfare-inferior to a then $\succ_g \notin C(\succ)$. Similarly $b \succ c \succ a$. Thus, we have two candidate welfare orderings: (i) $b \succ d \succ c \succ a$, which is not consistent

[3] See Manzini and Mariotti (2009) for a critical discussion of the Pareto approach.

[4] Bernheim and Rangel (2007, 2009) call the Pareto relation the unambiguous choice relation.

with $\succ_f$, and (ii) $b \succ c \succ d \succ a$, which is not consistent with $\succ_g$. The two welfare relations that are C-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.$\diamondsuit$

The problem of attaching welfare preferences to a behavioral dataset is related, though not identical, to the fundamental question of social choice theory: to formulate "social welfare" preferences that aggregate the preference relations of individuals in a society. In a typical social choice analysis, 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 in the sense that our aim is to attach a welfare relation to a profile of preference relations.[5] In social choice theory, each of the preference relations in a profile represents an individual in the society while in our approach each preference relation represents the same individual in a particular frame. The goal of social choice is to identify a society's welfare preferences while our goal is to uncover those of an individual.

We depart from standard social choice analysis in two major ways. First, we investigate potential cognitive deviations from an underlying welfare preference relation rather than the aggregation of autonomic and probably conflicting preference relations of different individuals. In a social choice context, this is analogous to the assumption that all preference relations in a profile are obtained from the same source relation according to a particular rule. Second, we study situations in which the data is a *set* of orderings rather than a *vector* of orderings. Thus, we do not specify which frame results in a particular preference relation and we do not account for whether the same preference relation appears in different frames. In the context of social choice, this is analogous to the combination of anonymity of individuals and invariance to the frequency of each preference relation in the population.

The recent literature on welfare analysis in the context of choice without framing effects (including Cherepanov, Feddersen and Sandroni (2008), Manzini and Mariotti (2008), and Masatlioglu, Nakajima and Ozbay (2009)) is also related to our analysis. In this literature, an observer records data on frame-independent choice behavior. The observer postulates that this data is the outcome of a specific procedure that uses an underlying preference relation in some structured way. Researchers in this literature identify conditions under which choice data is consistent with the postulated procedure and infer the underlying welfare preferences from choice data.[6] Green and Hojman (2007) advocate a different approach to welfare in the context of choice without framing effects. According to their approach, the decision maker has in mind several conflicting considerations that he "aggregates" in making choices. Given choice data, Green and Hojman characterize the set of the possible considerations that could have generated this data, and use this information in making welfare judgements.

In the next three sections, we demonstrate our approach using three scenarios. In the first, the

[5] Chambers and Hayashi (2009) investigate how to attach welfare relations to stochastic choice functions. Their analysis is similar to that of the multi-profile analysis in social choice theory in the sense that it imposes conditions that connect *across* stochastic choice functions the operation of attaching preference relations to choice data.

[6] In Cherepanov, Feddersen and Sandroni (2008), for example, an observer postulates that the individual follows a choice procedure that uses an underlying preference relation and a set of rationales. Given a set of alternatives, the decision maker identifies those alternatives that are maximal according to one of the rationales and then he chooses the most preferred alternative among them. The observer's goal is to identify the underlying preference relation from choice data.

decision maker satisfices and his aspiration goal is affected by framing as in the "busy individual" example. In the second, frames cause the individual to make small errors when evaluating the alternatives, as in the organization example. In the third, the frame highlights a single alternative, which benefits from a utility bonus in the spirit of the status quo bias. In each of these three scenarios, we begin by specifying the consistency function that describes the process by which the welfare preferences may be distorted by frames. We then identify conditions under which a behavioral dataset is consistent with the cognitive process behind the consistency function. When the dataset is consistent, we extrapolate from it the set of candidate welfare preference relations. The final section discusses possible extensions of our framework.

2. Satisficing

In this section, we have in mind a real-life dilemma. We observe a busy individual making choices from subsets of $\{a, b, c, d\}$. We know that whenever he makes a choice, the individual examines the available alternatives in the order a, b, c, d. We note that his behavior on Mondays is consistent with the preference relation $c \succ_m d \succ_m b \succ_m a$ and on Tuesdays with the preference relation $b \succ_t c \succ_t d \succ_t a$. How does the individual rank the alternatives b and c – by his preference on Mondays or by his preference on Tuesdays?

The following argument may make sense in some cases: When choosing from $\{a, b, c, d\}$ on Mondays, the decision maker considers a and b and finds them to be non-satisfactory. He then considers c and does find it to be satisfactory. Because on Mondays c is satisfactory and b is not, we conclude that c is welfare-superior to b. This conclusion is consistent with the decision maker's choice behavior on Tuesdays when he is probably busier and thus extends the set of satisfactory elements to include b.

The implicit premise in the above inference is that the decision maker follows a procedure of choice that Herbert Simon called Satisficing. A satisficer has in mind some aspiration goal, which induces a partition of the set of alternatives X to satisfactory and non-satisfactory elements according to his underlying welfare preferences. When making choices, he assesses the available alternatives in some *known* order O where aOb means that the decision maker examines a prior to b. He chooses the first satisfactory alternative he encounters and, if there are none, the last available alternative. For every aspiration goal, the decision maker's choices are consistent with maximizing a unique ranking of the elements in X.

Satisficing behavior may emerge when assessing the precise value of an alternative is costly while figuring out whether an alternative is simply "good enough" is less so. For example, in choosing among risky investment plans, it may be difficult to assess the expected return on a given plan, but not as difficult to assess whether the expected return exceeds the return obtained by a colleague. Satisficing may also emerge when there are search costs involved in considering an additional alternative, which the decision maker wishes to economize on. For example, interviewing an additional candidate for a job may be time consuming and thus a recruiter may settle on a candidate who is good enough.

In satisficing, the set of satisfactory alternatives may vary according to unobservable circumstances that are not related to the decision maker's welfare, i.e. what we refer to as frames. Going back to the above examples, an investor may use different colleagues as a source for his aspiration goal in different circumstances without us being able to observe this. Similarly, a recruiter may expand the set of satisfactory candidates on busier days; however, as observers, we do not know which days are which.

Let C be the consistency function that attaches to every preference relation $\succ$ all the possible rankings that might be displayed by a satisficer who examines the alternatives according to the order O and whose welfare preference relation is $\succ$. The set $C(\succ)$ includes all orderings that are obtained from $\succ$ by:

(1) partitioning the elements of X into two sets, S and $X \setminus S$, such that every element in S is $\succ$-superior to every element in $X \setminus S$,

(2) ranking the elements in S according to O and above all the elements in $X \setminus S$, and

(3) ranking the elements in $X \setminus S$ opposite to O.

A behavioral dataset Λ records different rankings displayed by a decision maker. In assessing the welfare of a decision maker whose behavior is recorded in Λ, we first test our conjecture that the decision maker is a satisficer that uses the order O. In other words, we check whether there is a preference relation $\succ$ such that $\Lambda \subseteq C(\succ)$. If this is indeed the case, we proceed to characterize the set of all preference relations that are C-consistent with Λ, i.e. preference relations that could have generated the behavioral dataset.

Given a behavioral dataset Λ, define the binary relation $\succ_R$ as follows:

$a \succ_R b$ if there are alternatives x and y such that [aOx and $a \succ_f x$] and [bOy and $y \succ_f b$] for some $\succ_f \in \Lambda$.

The logic behind this definition is that if $\succ_f$ is the outcome of satisficing behavior using the order O then aOx and $a \succ_f x$ imply that a is satisfactory and bOy and $y \succ_f b$ imply that b is not. Hence, a is welfare-superior to b. If, for example, a were not satisfactory, then either x is satisfactory, in which case x should have been ranked above a in $\succ_f$ or x is not satisfactory, in which case x should also have been ranked above a in $\succ_f$.

The following proposition uses $\succ_R$ to determine whether a given behavioral dataset can be explained by satisficing. It also establishes that when a behavioral dataset is C-consistent, any extension of $\succ_R$ to a complete and transitive relation is C-consistent with the dataset. In particular, $\succ_R$ is the maximal binary relation nested in any ordering consistent with the given dataset.

Proposition 1. For every behavioral dataset Λ:

(A) If $\succ_R$ is cyclic (where cyclic includes reflexive) then Λ is not C-consistent.

(B) If $\succ_R$ is acyclic then Λ is C-consistent. Moreover, any extension of $\succ_R$ to a complete and transitive binary relation is C-consistent with Λ.

Proof. We first note that if a preference relation $\succ$ is C-consistent with Λ then it nests $\succ_R$. Suppose $a \succ_R b$. Then there is a ranking $\succ_f \in \Lambda$ and two elements x and y such that aOx and $a \succ_f x$ and bOy and $y \succ_f b$. Because $\succ_f \in C(\succ)$, the element a is satisfactory and the element b is not implying that $a \succ b$. Part A immediately follows.

To prove part B, suppose that $\succ_R$ is acyclic. Consider an extension of $\succ_R$ to a complete, asymmetric and transitive relation $\succ$. To show that $\succ$ is C-consistent with the dataset Λ, we identify for every ranking $\succ_f \in \Lambda$ a set of alternatives $S(\succ_f)$ satisfying that (i) every element in $S(\succ_f)$ is both $\succ$-superior and $\succ_f$-superior to every element in $X \setminus S(\succ_f)$, (ii) the ranking of the elements in $S(\succ_f)$ is according to O and (iii) the ranking of the elements in $X \setminus S(\succ_f)$ is opposite to O.

Denote by m the O-minimal element in X. That is, m is the element that the decision maker considers last. For a ranking $\succ_f$, define $S(\succ_f) = \{a \mid a \succ_f m\}$. If there is an element $z \in S(\succ_f)$ such that $m \succ z$, then add m to $S(\succ_f)$. Thus every element in $S(\succ_f)$ is $\succ_f$-superior to every element in $X \setminus S(\succ_f)$.

Let us see that $\succ$ ranks every element $a \in S(\succ_f)$ above every element $b \in X \setminus S(\succ_f)$. If neither a nor b is m, then $a \succ_R b$ by setting $x = y =$

tical. When assessing the value of a given alternative, he may overestimate or underestimate its value. However, the assessment error is "small", i.e. it is less than the distance between two adjacent alternatives. Thus, evaluation errors change the ordering of two alternatives only when the alternatives are adjacent in the decision maker's underlying preference relation and the higher one is underestimated while the lower one is overestimated. In this case, we can infer the welfare ranking of two alteratives only if there is a third alternative between them. For example, if the dataset contains the two orderings $a \succ_f b \succ_f c$ and $b \succ_g c \succ_g a$ then we can conclude that the welfare preferences are $b \succ a \succ c$.

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

To examine whether small assessment errors can generate a given behavioral dataset Λ, we first define $a \succ_R b$ if there exists $\succ_f$ in Λ and an element x such that $a \succ_f x \succ_f b$. The following proposition uses $\succ_R$ to test whether a given behavioral dataset is C-consistent. It establishes that when the behavioral dataset is C-consistent, any extension of $\succ_R$ to a complete and transitive relation is C-consistent with the dataset. In particular, $\succ_R$ is the maximal binary relation nested in any relation that is C-consistent with the dataset.

Proposition 2. For every behavioral dataset Λ:

(A) If $\succ_R$ is cyclic then Λ is not C-consistent.

(B) If $\succ_R$ is 3-acyclic then Λ is C-consistent.[7] Moreover, any extension of $\succ_R$ to a complete and transitive binary relation is C-consistent with Λ.

Proof of 2.A. We show that if there exists a preference relation $\succ$ that is C-consistent with Λ then $\succ$ nests $\succ_R$ and therefore $\succ_R$ is acyclic. Suppose $a \succ_R b$. Then, there is $\succ_f \in \Lambda$ and an element x such that $a \succ_f x \succ_f b$. Assume to the contrary that $b \succ a$. Then, because $a \succ_f b$, there are no elements ranked between b and a in $\succ$. Therefore, either x is ranked above b in $\succ$ contradicting $a \succ_f x$ or x is ranked below a in $\succ$ contradicting $x \succ_f b$. Thus, we cannot have that $b \succ a$, and because $\succ$ is complete we obtain that $a \succ b$. We conclude that $\succ$ nests $\succ_R$.

Proof of 2.B. We first show that since $\succ_R$ is 3-acyclic, it is acyclic. Suppose $\succ_R$ has a cycle and consider the *shortest* one $x_1 \succ_R x_2 \succ_R ... \succ_R x_K \succ_R x_1$. Since $\succ_R$ is 3-acyclic, we have that $K > 3$. Because $x_1 \succ_R x_2$, there exist $\succ_f$ in Λ such that $x_1 \succ_f x \succ_f x_2$. For $k = 3$ or $k = 4$, the element x is not equal to x_k. If $x_k \succ_f x$ then by definition $x_k \succ_R x_2$, and if $x \succ_f x_k$ then $x_1 \succ_R x_k$. In either case we have a shorter cycle.

The relation $\succ_R$ is therefore acyclic and can be extended to a complete and transitive relation $\succ$. Assume to the contrary that $\succ$ is not C-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 happen because (i) if $x \succ_f b \succ_f a$

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

then $x \succ_R a$ contradicting $a \succ x$, (ii) if $b \succ_f x \succ_f a$ then $b \succ_R a$ contradicting $a \succ b$, and (iii) if $b \succ_f a \succ_f x$ then $b \succ_R x$ contradicting $x \succ b$. ■

There may be more than one preference relation that is C-consistent with a given dataset. For example, if the behavioral dataset contains only one preference relation, then any ranking of the alternatives obtained from that relation by disjoint switches of adjacent elements is C-consistent with the dataset. Proposition 3 identifies a simple condition on a dataset that is necessary and sufficient for the existence of a unique preference relation that is C-consistent with the dataset.

Proposition 3. Assume a behavioral dataset Λ is C-consistent. There exists a unique preference relation that is C-consistent with Λ if and only if for every two elements a and b at least one of the following holds:

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

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

Proof. To prove the if part, it is enough to show that the transitive closure of $\succ_R$ relates every two alternatives a and b because by proposition 2 the relation $\succ_R$ and hence its transitive closure are nested in any relation consistent with Λ.

If (i) holds for two alternatives a and b then $\succ_R$ relates a and b. If (ii) holds for a and b, then we have two rankings 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 $a \succ_g b$; otherwise we would have that $x \succ_f a \succ_f b$ and $b \succ_g a \succ_g x$ implying both $x \succ_R b$ and $b \succ_R x$. Thus, we have $x \succ_f a \succ_f b$ and $a \succ_g b \succ_g x$. Therefore, $x \succ_R b$ and $a \succ_R x$ and the closure of $\succ_R$ connects a and b.

To prove the only if part, consider two alternatives a and b such that both (i) and (ii) fail. Let U (D) denote the set of elements that are above (below) both a and b in all the rankings in Λ. Since (i) and (ii) fail, the sets U and D contain all the elements of X other than a and b. We now show that the transitive closure of $\succ_R$ does not relate a and b. Hence by Proposition 2 both rankings of these two alternatives are possible. Assume to the contrary that $a \succ_R x_1 \succ_R x_2 \succ_R \ldots \succ_R x_k \succ_R b$. Then $x_k \in U$ because $x_k \succ_R b$ implies there is a ranking $\succ_f$ and an element y such that $x_k \succ_f y \succ_f b$. By iterating this argument, we obtain that $x_1 \in U$. But by $a \succ_R x_1$, we have that x_1 is ranked two or more places below a in some ranking and thus $x_1 \in D$.■

4. Highlighting a single alternative

We survey an individual repeatedly to learn about his tastes and note that his reports are inconsistent. We suspect that in each report the individual assigns a utility bonus to one of the

alternatives while preserving the ranking of the others. This may be because the decision maker recently chose that alternative or because it reflects a current consumer trend. In other words, the high ranking of an alternative may be an outcome of a psychological "status quo bias". We do not know which alternative gets the bonus nor its size and wish to elicit the individual's preferences net of the psychological bias.

Consider, for example, a dataset containing the two relations $a \succ_f b \succ_f c$ and $b \succ_g c \succ_g a$. In this case, we can conclude that either a or c must be ranked last in any welfare preferences that are consistent with the dataset. If c is ranked last, then c got a bonus in g and thus the welfare preferences must be $b \succ a \succ c$, which is consistent with b getting a bonus in f. If a is ranked last then a got a bonus in f and the welfare preferences must be $b \succ c \succ a$, which is consistent with no element getting a meaningful bonus in g. In any case, we can conclude that $b \succ a, c$.

More formally, let C be the consistency function that assigns to a preference relation $\succ$ all the relations that are obtained from $\succ$ by advancing a single alternative to a weakly higher position, i.e. $C(\succ) = \{\succ_f \mid$ there is at most one element b such that $a \succ b$ and $b \succ_f a$ for some element $a\}$. The set $C(\succ)$ contains $\binom{|X|}{2} + 1$ preference relations including the relation $\succ$.

An observer conjectures that a given dataset Λ was generated by the above mechanism. Clearly, if Λ contains a single ordering $\succ_f$, then $\succ_f$ and every preference relation that is obtained from it by moving down a single alternative is C-consistent with Λ. We now discuss a simple way of identifying whether a dataset Λ containing two or more orderings is C-consistent, and eliciting an ordering that is C-consistent with Λ.

Denote by $T \subset X$ the largest set of elements such that all orderings in Λ rank the elements in T below the elements in $X \setminus T$ and agree on the ordering of the elements in T. Thus, T is the common *Tail* of all the orderings in Λ. The dataset Λ consists of more than one ordering and hence $T \neq X$. Of course, T may be empty as in the above example.

Let $A \subseteq X \setminus T$ be the set of elements such that $x \in A$ if x appears right above the elements of T in some $\succ_f \in \Lambda$. That is, $A = \{x \mid x$ is $\succ_f$ -minimal in $X \setminus T$ for some $\succ_f \in \Lambda\}$. Denote by $\Lambda_x \subseteq \Lambda$ the set of all orderings in which the alternative x is minimal among all the elements in $X \setminus T$. In the above example, $A = \{a, c\}$, $\Lambda_a = \{\succ_g\}$ and $\Lambda_c = \{\succ_f\}$.

Proposition 4. If a dataset Λ contains at least two orderings and is C-consistent, then

(i) the set A contains two alternatives, and

(ii) for one of the alternatives $b \in A$, the ordering of $X \setminus \{b\}$ is identical across all orderings in Λ_a where a is the other element in A and expanding this ordering by positioning b right below a is C-consistent with Λ.

Proof. (i) Assume $\succ$ is C-consistent with Λ. By definition, $|A| \geq 2$. Assume to the contrary that $|A| > 2$. Then there are three elements a, b and c and three orderings $\succ_f$, $\succ_g$ and $\succ_h$ in Λ such that

$a, b \succ_f c$, $a, c \succ_g b$ and $b, c \succ_h a$. Suppose without loss of generality that a is $\succ$-minimal among a, b and c. Then, a jumped in $\succ_f$ and $\succ_g$ implying that the relative ranking of b and c in these two rankings should be identical in contradiction to $b \succ_f c$ and $c \succ_g b$.

(ii) Assume that $\succ$ is C-consistent with Λ. Let a and b be the two alternatives in A such that $a \succ b$. Then the element b jumps in all the orderings in Λ_a and hence any ordering in Λ_a restricted to $X \setminus \{b\}$ is identical to $\succ$ restricted to the same set. If b is $\succ$-superior to all the elements in T, then $\succ$ is the ordering described in (ii). Otherwise, there is an element $x \in T$ such that $x \succ b$. This implies that b jumps in all the orderings in Λ and they all rank b above the elements of T implying that the ordering described in (ii) is C-consistent with Λ. ■

Thus, in order to test whether a dataset Λ is C-consistent, we need only to ascertain whether one of the two candidate orderings described in Proposition 4 is C-consistent with Λ. In addition, the proof of Proposition 4 implies that if an ordering $\succ$ is C-consistent with Λ then $\succ$ is obtained from one of the two candidate orderings described in the proposition, $\succ'$, by moving down in $\succ'$ the $\succ'$-minimal element in $X \setminus T$.

5. Discussion

This paper analyzes situations in which the same decision maker displays different preference relations in various circumstances that differ in payoff-irrelevant parameters. An observer conjectures that this is the result of systematic deviations from a preference relation that reflects the individual's welfare and he wishes to elicit the welfare relation from the observed preference relations. In the previous sections, we demonstrated the elicitation process for several scenarios. We conclude with a discussion of possible modifications of the framework.

5.1. Observing the frame

In our framework, an observer knows which preference relations could be obtained from the welfare preferences. This is summarized in the consistency function C. However, the observer does not know what the actual frames are or the way in which each specific frame distorts the underlying preferences.

An alternative framework would be one in which the observer knows these details. Formally, a behavioral dataset is a set $\Lambda = \{(\succ_f, f)\}$ in which each observation includes a preference relation and the distorting frame. The way in which each frame distorts the welfare preferences is summarized by a consistency function that takes both the underlying preference relation and the frame as parameters. We denote by $C(\succ, f)$ the set of behavioral preference relations that can be obtained from the underlying welfare ordering $\succ$ under the frame f. A relation $\succ$ is C-consistent with Λ if for every $(\succ_f, f) \in \Lambda$ we have that $\succ_f \in C(\succ, f)$. The following example demonstrates the elicitation process in the modified framework.

Highlighting multiple alternatives. Consider a case in which a set of alternatives is highlighted by some external mechanism. Examples include a web site where some of the alternatives are

presented in bold font or a grocery store that positions some products near the cashier in order to attract attention to them. A highlighted alternative gets a non-negative utility bonus and its ranking with respect to non-highlighted alternatives is weakly improved. The ranking among the non-highlighted alternatives is identical to that of the welfare preference relation while the ranking among the highlighted alternatives may differ from that of the welfare relation. Formally, the set of highlighted alternatives f is a subset of X and $C(\succ, f) = \{\succ_f \mid \text{if } [b \succ a \text{ and } a \succ_f b] \text{ then } a \in f\}$.

Suppose we observe *both* the preference relation $\succ_f$ and the set of highlighted elements f. Then we can infer the welfare ranking between a non-highlighted alternative and any element that is $\succ_f$-inferior to it. Define $a \succ_R b$ if there exists a frame f such that $a \succ_f b$ and $a \notin f$. We now show that for any dataset $\Lambda = \{(\succ_f, f)\}$ the set of orderings that are C-consistent with Λ is the set of all complete and transitive extensions of $\succ_R$.

Let $\succ$ be a preference relation that is C-consistent with Λ. Note that if $a \succ_R b$ then in some frame f, we have that $a \notin f$ and $a \succ_f b$. This implies that $a \succ b$ since otherwise $b \succ a$ and $a \succ_f b$ would imply that $a \in f$. Thus, $\succ_R$ is nested in any preference relation that is C-consistent with Λ. Consider a complete and transitive extension of $\succ_R$ denoted by $\succ$. In order to determine whether $\succ$ is C-consistent with Λ, we need to verify that if $a \succ b$ and $b \succ_f a$ then $b \in f$. Otherwise, $b \notin f$ and $b \succ_f a$ imply that $b \succ_R a$ and thus, since $\succ$ extends $\succ_R$, we cannot have that $a \succ b$. $\diamondsuit$

5.2. Cardinal Utilities

In our framework, welfare preferences are ordinal. There are cases, however, in which describing the cognitive process that distorts welfare preferences requires a notion of the intensity of these preferences. For example, in the context of advertising, the decision maker may prefer product a to product b; however, the number of times he views the same advertisement for each product may influence his choice between them. In order to describe the magnitude of the advertising bias in the decision maker's preferences, we need to add a notion of cardinal utility.

Advertising. The decision maker is characterized by a utility function u that assigns positive values to different alternatives and represents his welfare preferences. Before making a choice, the decision maker views advertisements for the various alternatives. We refer to this information as a frame and define it formally as a function $i : X \to N$ that assigns to every alternative $x \in X$ the number of ads $i(x)$ for that alternative. After viewing the ads, the decision maker maximizes $i(x)u(x)$ rather than $u(x)$. We can observe i and the resulting preferences $\succ_i$ but not u.

We say that $\succ$ is consistent with the dataset $\Lambda = \{(\succ_i, i)\}$ if there is a utility representation u of $\succ$ such that for every observation $(\succ_i, i)$ the function $u(x)i(x)$ represents $\succ_i$. The existence of such a function u is equivalent to the existence of a solution to a system of inequalities in the $|X|$ unknowns $\{u(x)\}_{x \in X}$, where each inequality is of the form $i(x)u(x) > i(y)u(y)$ for $x \succ_i y$. In particular we can conclude that $x \succ y$ if $i(y) \geq i(x)$ and $x \succ_i y$ for some $(\succ_i, i)$ in the dataset. $\diamondsuit$

References

Bernheim, B. Douglas, and Antonio Rangel (2007). Toward Choice-Theoretic Foundations for Behavioral Welfare Economics. *American Economic Review Papers and Proceedings*, 97(2), 464-470.

Bernheim, B. Douglas, and Antonio Rangel (2009). Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics. *Quarterly Journal of Economics*, 124(1), 51-104.

Chambers, Christopher P. and Takashi Hayashi (2008). Choice and Individual Welfare. mimeo.

Cherepanov, Vadim, Timothy Feddersen and Alvaro Sandroni (2008). Rationalization. Working paper, Kellogg School of Management.

Green, Jerry R. and Daniel A. Hojman (2007). Choice, Rationality and Welfare Measurement. Harvard Institute of Economic Research Discussion Paper No. 2144 and KSG Working Paper No. RWP07-054.

Manzini, Paola and Marco Mariotti (2008). Categorize Then Choose: Boundedly Rational Choice and Welfare. mimeo.

Manzini, Paola and Marco Mariotti (2009). Choice Based Welfare Economics for Boundedly Rational Agents. mimeo.

Masatlioglu, Yusufcan, Daisuke Nakajima and Erkut Y. Ozbay (2009). Revealed Attention. mimeo.

Salant, Yuval and Ariel Rubinstein (2008). (A, f): Choice with Frames. *Review of Economic Studies*, 75(4), 1287-1296.