Behavioral Welfare Analysis and Revealed Preference: Theory and Experimental Evidence

JOB MARKET PAPER

Daniele Caliari*

April 9, 2020

Abstract

There is no general consensus on how welfare analysis should be carried out for individuals that violate the Weak Axiom of Revealed Preference. Some proposed solutions ignore data where violations occur, not accounting for possibly important pieces of information. We study procedures that elicit welfare relation from dataset, denoted as *Welfare Methods*. We adopt a model-free approach and propose a series of normative principles. In particular, we propose a property called Informational Responsiveness. It states that a welfare method that ranks A and B should not ignore relevant observations; namely those where either A or B is chosen and both are available. In our main theoretical results we show the relevance of Informational Responsiveness (Proposition 1) and we characterize a method that counts revealed preference relations (Theorem 2).

We test the joint importance of Informational Responsiveness and Revealed Preference using experimental data. We conduct a novel experiment in which subjects firstly face a sequence of questions regarding time and risk outcomes and secondly report the welfare relation over some of the alternatives. We find that welfare methods that violate Informational Responsiveness have significantly worse performances in terms of both identification of the reported best element and the entire welfare relation (Table 3).

Keywords: Welfare analysis, Bounded rationality, Stochastic choice, Revealed preference.

*Department of Economics and Finance, Queen Mary University of London. E-mail: d.caliari@qmul.ac.uk
I am indebted to Marco Mariotti and Christopher Tyson for the numerous advices. I also thank Sean Horan, Ivan Soraperra, Georgios Gerasimou, David Freeman, Maria Vittoria Levati, Aniol Llorente-Saguer and seminar participants at Queen Mary University, CEPET Workshop, RES 2019 Annual Conference, EEA-ESEM 2019 Summer Congress.

1 Introduction

This paper is concerned with the study of violations of the Weak Axiom of Revealed Preference[1] and of how welfare analysis should be performed. If no violations are observed, welfare analysis is trivial and the elicited preference relation is equivalent to the maximized one. However, overwhelming evidence has been produced in both psychology and economics to show that individuals often do not behave according to standard assumptions of rationality.[2] Not only, but the literature has proposed several models, often mutually exclusive, to explain the same cognitive constraint.[3] Each model provides a different way to construct a revealed preference relation. Therefore, it is still an open question how welfare analysis, which is normally guided by a well-defined preference relation, has to be performed in these cases.

We study welfare methods as family of maps that associate binary relations to behavioural types.[4] Few attempts have been made to study appealing properties of these maps. Our first aim is to provide normative principles that can be guidelines for welfare analysis when standard revealed preference does not apply. Importantly, all the properties can be strengthen or weakened without losing their normative interpretation.[5]

First and foremost, we propose the following normative principle: a method that ranks two alternatives A and B must use all the relevant feasible evidence about A and B. We call this condition Informational Responsiveness.[6] Its formalization exploits the potential pivotal role of "important" (the term refers to situations where A is chosen and B is available or vice versa) observations in the case of two alternatives being judged as indifferent.[7] Formally, if A is indifferent to B, more evidence in favour of A should turn the judgement in its favour. A violation would imply that the welfare method is ignoring that particular evidence. The foundation of this requirement lays in the idea that more information must bring to finer conclusions. In the

[1] WARP's definition is as follows: if an alternative **x** is chosen when **y** is available then **y** is not chosen when **x** is available. In general, Weak and Strong Axiom of Revealed Preference are not equivalent. Sen (1971) proved the equivalence under certain conditions. Our experiment does not always meet these conditions, however a deeper study of the Strong Axiom of Reveled Preference found no further information as presented in the Online Appendix. Therefore, we focus on the simpler weak version.

[2] Violations regard not only WARP (Echenique et al., 2011) but also stochastic properties such as independence from irrelevant alternatives (Tversky & Russo, 1969), weak stochastic transitivity (Tversky, 1969) and regularity (Huber et al., 1982), (Iyengar & Kamenica, 2010).

[3] The existence of different models that explain similar situations regards, for instance, how individuals deal with complex choice problem, in particular when the number of alternatives is high. Both in deterministic and stochastic literature two main lines of models have been developed: (i) (degenerate) attention models has been developed among many by Masatlioglu et al. (2012), Lleras et al. (2017), Manzini & Mariotti (2014a), Echenique et al. (2018), Cattaneo et al. (2018); (ii) (uniform) attention models by Frick (2016), Fudenberg et al. (2015).

[4] In choice theory literature, individuals are often identified by their choices. Hence, those who make the same choices are defined as to be of the same behavioral type.

[5] This feature is in line with the idea that an axiomatization should be satisfactory. In Krantz et al. (1971): "One demand is for the axioms to have a direct an easily understood meaning in terms of empirical operations, so simple that either they are evidently empirically true on intuitive grounds or it is evident how systematically to test them."

[6] To the best of our knowledge this property has been firstly introduced in voting theory by Goodin & List (2006) under the denomination of "One Vote Responsiveness".

[7] In demand theory an analogous axiom is local non-satiation as it rules out "thick" indifference curves.

literature of preference elicitation this principle has been highlighted by Rubinstein & Salant (2012) when, commenting the Pareto approach proposed by Bernheim & Rangel (2009), they wrote: "the resulting Pareto relation is typically a coarse binary relation that becomes even more so as the behavioural data set grows."

We argue that the necessity of this axiom is related to its weakness, non-triviality, and relevance. We show that structurally different welfare methods satisfy it (*weakness*) but some do not (*non-triviality*). In such cases, we show that the violation can potentially lead to paradoxical results (*relevance*). The following remark exposes with a simple example this latter argument in favour of the necessity of Informational Responsiveness:

Remark 1. Suppose a dataset contains multiple observations over the binary comparison $\{x, y\}$ where x is chosen 99 times while y only once. The conclusion that x and y are either equally valuable or incomparable would be paradoxical. We show that this result is due solely to the violation of Informational Responsiveness.

Nonetheless, there are situations where Informational Responsiveness may not be optimal. The following two remarks provide a brief theoretical and empirical discussion.

Remark 2. Sen (1971) and Arrow (1959) propose two main approaches to revealed preference: one based on all possible subsets and one based only on binary sets. The former is the most common[8] and it is in line with Informational Responsiveness. However, the latter has also received attention in the literature (Manzini & Mariotti, 2012) and it constitutes a theoretical example of violation of our proposed requirement.

Remark 3. Iyengar & Kamenica (2010) propose an experimental setting where they investigate choices in sets of 3 and 11 lotteries. They observe a violation of regularity,[9] with the simplest alternative (degenerate lottery) chosen only 16% of times in the small set and 63% of times in the big set. This is just one of the many empirical examples of choice reversal that raises doubts regarding which sets should be considered when performing welfare analysis.

Theoretically our work locates in the axiomatic approach recently proposed by Nishimura (2017) and Horan & Sprumont (2016). However, it differs from both. Unlike the former our primitives are choice observations and not preference relations and unlike the latter we propose normative principles that deals with the problem of informational collection (informational responsiveness) and arrangement (revealed preference approach).

We provide two main theoretical results:

[8] Sen (1971) lists three methods, however two his proposal are equivalent under the assumption of element-valued choice function. The standard case is defined as: (i) xRy if and only if for some S, $x \in c(S)$ and $y \in S$; (ii) $x\bar{R}y$ if and only if $x \in c(\{x, y\})$.

[9] Regularity is a well-known necessary condition for Random Utility Models. It has been firstly introduced by Marschak & Block (1960) and it is defined as $p(x, S) \geqslant p(x, T)$ when $S \subseteq T$.

- We show that Informational Responsiveness is the key axiom that gives rise to a class of methods that can infer the underlying deterministic utility of a variety of stochastic models among which, for instance, i.i.d. Random Utility Model (Proposition 1 and 2);[10]

- We characterize both the simple and the revealed preference counting procedures (Theorem 1 and 2). In the latter case, we show that the procedure has two appealing properties: (i) in certain cases it is equivalent to the Minimum Swaps Methods (Proposition 3) proposed by Apesteguia & Ballester (2015); (ii) it can be used as foundation for other important methods such as Graph Centrality (e.g. Eigenvector Centrality) methods and the Transitive Core method proposed by Nishimura (2017).

In the second part of the paper we test our theory using new experimental data. We answer the following two questions that constitute together the premise and the testing of our theoretical analysis.

- Premise: Do individuals consistently reveal welfare in different choice problems, e.g. in time or risk preferences, with attraction effect or choice overload?

- Test: If not, how should we measure welfare when individuals violate the Weak Axiom of Revealed Preference? Particularly: is *Informational Responsiveness* effective in discriminating welfare methods? And how important are revealed preference relations?

To the best of our knowledge, our experimental design is a novelty. We use a choice elicitation design divided in two parts: Time and Risk. Subjects are asked to choose among delayed payment plans and lotteries. They are paid for one random decision for each part.[11] As in Manzini & Mariotti (2010) we collect the entire choice function[12] regarding four alternatives, that we call MAIN alternatives. This is a crucial element for two reasons: (1) full observability is usually a necessary requirement for testing axioms of choice;[13] (2) it guarantees that part of the dataset is completely symmetric with respect to the MAIN alternatives. The remaining questions are either neutral or they contain asymmetric dominance[14] and choice overload problems.[15] The reasoning behind this structure is to test if information contained in questions that are potentially doomed by behavioural effects can be important to define the welfare relation of the subjects.

[10]The result can be easily generalized to the family of stochastic choice models that satisfy the property of Acyclicity or Item Acyclicity proposed by Fudenberg et al. (2015).

[11]This payment structure is standard; see Hey & Carbone (1995), Agranov & Ortoleva (2017).

[12]Namely, we collect answers about all the non-empty subsets with more than two elements. In our case there are six binary sets, four ternary sets and one quaternary set.

[13]For instance, the result [WARP $\Leftrightarrow$ SARP] relies on the full observability of all non-empty subsets, Sen (1971).

[14]The behavioural effect known as asymmetric dominance deals with ternary sets where one alternative is clearly dominated by one of the other and while the remaining ones have similar value. A typical observation in this environment is attraction effect, see Huber et al. (1982) and Natenzon (2019).

[15]With choice overload we intend a situation where the number of alternatives in a choice set makes difficult for the decision maker the evaluation of all of them. This effect has been investigated empirically by Iyengar & Kamenica (2010) and theoretically, among many, by Masatlioglu et al. (2012), Lleras et al. (2017), Frick (2016)

Dataset asymmetry represents a challenge to the answer of our first question. The comparison of consistency of choice among different and non-symmetric parts of the dataset is a well-known problem.[16] We address this problem developing a new index of rationality that is robust to the structure of the dataset. We make use of the perturbation parameter of the logit model to match via Monte Carlo simulation the average number of WARP violations in each part of the dataset. Although the strong assumptions and limitations of the index our evidence shows that there is a connection between preference revelation and the index of rationality.

Our main objective (second question) is to evaluate welfare methods in view of our theoretical results. A the end of the experiment we ask subjects, in a non-incentived way, to rank the four MAIN alternatives. We consider this relation as a benchmark for evaluating how welfare methods perform on the dataset. The reliability of the reported welfare relation is empirically strong.[17] Our main findings are the following.

First, we find that a good proportion of subjects never violate WARP in time preferences (37%). Conversely, and in line with the literature (Agranov & Ortoleva, 2017), almost no subjects satisfy WARP in risk preferences (6%). The average number of violations of WARP reflects this finding: the average in time is 11.26 while in risk is 24.65 (the difference is significant with $p \approx 0$), and robust if we focus only on subjects that violated WARP at least once.

Second, we observe that methods that satisfy Informational Responsiveness outperform the other welfare methods. A relevant example of this latter is the method proposed by Bernheim & Rangel (2009). When asked to uniquely identify the best reported alternative this method is outperformed by 30% in time and 50% in risk.[18] When limited to a set identification exercise, more in line with its conservative approach, it is still outperformed by 15% in time and 20% in risk. These results are robust when we limit ourselves to the non-empty subsets of MAIN alternatives. Similarly, when asked to uniquely identify the entire welfare relation, Bernheim & Rangel method is outperformed by 20% in time and 25% in risk preferences.

Third, we compare the identification power of the simple counting, that satisfy a stronger version of Informational Responsiveness, with the counting revealed preference procedure. We find that it is outperformed by 6% in time and 4% in risk. This suggests on one hand that our property is not sufficient and that a stronger version could have negative effects; on the other hand that revealed preferences play an important role in the identification process.

Four, we analyse these results using a measure of completeness for models developed by

[16] See Andreoni et al. (2013) for a survey of the literature.

[17] The reliability of the reported ranking is confirmed by the following statistics: in time preferences 69 out of 70 rational subjects reported the correct optimal alternative and 61 out of 70 reported correctly the entire welfare relation. This statistic is repeated in risk preferences with respectively 10 out of 12 subjects reporting the correct optimal alternative and 9 out of 12 the correct welfare relation. Two subjects reported the opposite ranking to the one they rationally employed in their choices. This probable mistake does not affect our results since every method will clearly fail to identify these subjects.

[18] These percentages are calculated on the total number of subjects. For example, in Time the method proposed by Bernheim & Rangel (2009) uniquely identifies the correct best alternative of 59% of the subjects while the counting revealed preference method of 87%.

Fudenberg et al. (2019). The main advantage of this measure is to provide a power of methods with respect to the most naive and most sophisticated method. The idea is as follows: subjects that are not identified by the most sophisticated method are considered an irreducible error; while subjects that are identified by the most naive method are considered as trivial. All the methods analysed are ranked using not the total proportion of identified subjects but only the proportion of non-trivial and "feasible to be identified" subjects. We confirm that methods that satisfy Informational Responsiveness and are based on standard revealed preference are significantly more complete.

Five, we directly test Informational Responsiveness. We apply an optimal weighting algorithm on the dataset in order to maximize the identification process. We find that, when asked to maximize a combination of the best reported element and the entire welfare relation, the algorithm gives strictly positive weights to all sets with only one exception (negative weights) happening in time preferences for sets potentially doomed by asymmetric dominance.

1.1 Related Literature

The theoretical part of the paper is related to the small literature on welfare methods: Green & Hojman (2007), Salant & Rubinstein (2008), Bernheim & Rangel (2009), Rubinstein & Salant (2012), Manzini & Mariotti (2014b), Apesteguia & Ballester (2015), Horan & Sprumont (2016), Nishimura (2017). We also contribute to order theory through the characterization of counting procedures for datasets with multiple observations and missing data is a novelty. This result is connected with two axiomatizations of counting procedures in tournaments (Rubinstein, 1980) and directed graph (van den Brink & Gilles, 2003). Finally, our results can be applied through the law of large numbers to a variety of stochastic choice models including: i.i.d. Random Utility models (Marschak & Block, 1960), Luce model and Additive Perturbed Utility models (Fudenberg et al., 2015).

The index of rationality based on the perturbation of a data generating process such as the logit model is connected with the literature on rationality indexes and power measures. The most prominent example is the Selten measure (Selten, 1991) of which a special case is the Bronars hypothesis (Bronars, 1987). Our index, being robust to the dataset structure, overcomes a problem common to other indexes such as Afriat's index (Afriat, 1973), minimum number of observations to remove to rationalize the data (Houtman & Maks, 1985), number of violations of consistency axioms (Swofford & Whitney, 1987) and (Famulari, 1995), minimum number of swaps (Apesteguia & Ballester, 2015).

The experimental part firslty related to the few choice elicitations experiments such as Manzini & Mariotti (2010) and Barberá & Neme (2017). Secondly it is related to the literature on stochastic choice and choice deferral. Our design shares some features with existent experiments. Nonetheless none of the following papers have focused on welfare analysis and therefore all of them have key differences with ours. Some are restricted to binary comparisons: Agranov &

Ortoleva (2017), Hey & Carbone (1995), Danan & Ziegelmeyer (2006), Hey (2001), Cavagnaro & Davis-Stober (2014), Sopher & Narramore (2000), Chabris et al. (2009). Others collect data only on particular sets: Harbarugh et al. (2001) elicited choices from 11 different sets with cardinality from 3 to 7; Iyengar & Kamenica (2010) elicited choices from sets of either 3 or 11 gambles; Haynes (2009) collected response times but he elicited choices only from sets of either 3 or 10 prizes; Iyengar & Lepper (2000) elicited choices from sets of either 6, 24 or 30 alternatives; Sippel (1997) elicited 10 choices from budget sets regarding 8 alternatives.

1.2 Structure of the paper

The paper's structure is as follows: in Section 2 we introduce the framework and present the theoretical results. We also describe the welfare methods that will be analysed subsequently. In Section 3 we present in details the experimental design. The main experimental results as well as the index of rationality are presented in Section 4. All of them are divided with respect to time and risk preferences. The Appendix contains details regarding proofs and independence of the axioms. More details regarding the experimental design such as: parametrization of the alternatives, orders of the questions and questionnaire are contained in the Online Appendix.

2 Theory

Let X be a finite set of alternatives and $\mathcal{X}$ the set of all non-empty subsets of X. Denote O as the set of all possible pairs (x, A) where $A \subseteq X$ and $x \in A$. A dataset D assigns a non-negative integer to each pair; we write $D(x, A) = 1$ to say that x has been chosen from A one time.[19] We denote $\mathcal{D}$ as the set of all possible datasets.

Denote as $\mathcal{R}(X)$ the set of all complete[20] and reflexive binary relations on X. A welfare method is a function $f : \mathcal{D} \to \mathcal{R}(X)$ that maps each dataset into a welfare relation. Welfare methods will be the objects of our analysis.

We denote $xR_f^D y$ to say "x is weakly better than y on the dataset D by a welfare method f". As an abuse of notation we write $xR_f^{D+(x,A)} y$ to define the weak preference over a dataset D to which we have added an observation where x is chosen from A.

It is useful to define two counting measures. The simple counting, denoted C_x, and the counting revealed preference relations, denoted C_{xy}.

$$C_x = \sum_{A \subseteq X} D(x, A) \qquad\qquad C_{xy} = \sum_{A \ni x,y} D(x, A)$$

[19] A dataset D can be seen as the frequency version of a stochastic choice function.

[20] Completeness is defined as follows: for all $D \in \mathcal{D}$, for all $x, y \in X$, either $xR_f^D y$ or $yR_f^D x$. Even if not directly stated as an axiom, Completeness plays a crucial role in all proofs and it is assumed throughout all the paper.

The counting choice method **CC** is defined as follows:

$$xR^D_{\mathbf{CC}}y \text{ if and only if } C_x \geqslant C_y$$

So far we haven't assumed neither acyclicity nor transitivity.[21] The reason is that this allows us to define the counting procedure applied on standard revealed preference relation as a welfare method denoted as **CRP**. The reader may note that generally if a generic welfare relation P is cyclic then in some sets it has no maximal elements. Its inclusion is driven by the following arguments: (1) **CRP** is the foundation for other important methods; (2) the acyclicity of $P^D_{\mathbf{CRP}}$ can itself be empirically tested and if the condition holds **CRP** can be used effectively as welfare method.

The counting revealed preference method **CRP** is defined as follows:

$$xR^D_{\mathbf{CRP}}y \text{ if and only if } C_{xy} \geqslant C_{yx}$$

2.1 Informational Responsiveness

One feature of welfare methods we want to capture is that they should use all relevant data to discriminate between x, y. As relevant we intend all data where x, y are observed by the decision maker and one of them is chosen. In order to formally state this idea we split in two parts a well-known condition known as Positive Responsiveness. This assumption is stated as follows: "if x is weakly better than y [xRy] and we observe x chosen from one more choice set then x becomes strictly better than y [xPy]".[22] This axiom is strong for two reasons: (1) the antecedent is concerned with both I and P (respectively the symmetric and asymmetric part of R); (2) x can be chosen from any possible choice set.

We weaken this axiom allowing choices to be only weakly positive signals of welfare and limiting the welfare relevant sets to those where x is chosen and y is available. Following these considerations we define two axioms. Note that all axioms we state hold for all $A \subseteq X$, $D \in \mathcal{D}$ and for all $x, y \in X$.

Axiom 1 (Informational Responsiveness **[IR]**[23])**.**

$$xI^D_f y \quad \& \quad x, y \in A \quad \Rightarrow \quad xP_f^{D+(x,A)}y$$

This is the main axiom of the paper; and the one that we test in our experiment. In the next section we argue that this axiom is a necessary condition for the function f. The violation of

[21] A binary relation P is acyclic if there exists no sequence of elements $(x_i)_{i=1}^n$ such that $x_1Px_2P\ldots x_nPx_1$. A binary relation P is transitive if for all $x, y, z \in X$, xPy and yPz imply xPz.

[22] May's Theorem contains exactly this formulation of the axiom (May, 1952).

[23] The consequent of this axiom is technically incomplete. We should define it when we both add and remove observations. The complete version is $xP^{D+(A,x)}y$ and $yP^{D-(A,x)}x$.

this condition implies that a method f does not consider the observation (x, A) as relevant for welfare.

Axiom 2 (Choice non-negativeness **[CNN]**).

$$xI_f^D y \;\Rightarrow\; xR_f^{D+(x,A)} y \qquad \& \qquad xP_f^D y \;\Rightarrow\; xP_f^{D+(x,A)} y$$

Notice that this axiom is satisfied when choice observations are summed and weighted but all are weakly positive signal of the welfare of chosen elements.

Finally, let $\Pi(X)$ be the set of all the permutations $\pi : X \to X$. For all π:

Axiom 3 (Neutrality **[NEU]**).

$$xR_f^D y \;\Leftrightarrow\; \pi(x) R_f^{\pi(D)} \pi(y)$$

This axiom is standard in the literature and it asserts that a welfare method cannot, a priori, favour or punish some alternatives over others.[24] Since our theory does not rely on any additional information about alternatives or models, Neutrality seems to be a reasonable assumption.

2.2 Weakness, Non-Triviality and Relevance of IR

We claim that IR should be considered a necessary condition for welfare methods due to three features: Weakness, Non-Triviality and Relevance.[25] Weakness depends on the fact that the antecedent constrains the mapping only at the indifference. Non-Triviality is stated as follows: "there exist some welfare methods proposed by the literature that violate IR". In Figure 1, at the end of this section, we show how CNN and NEU are trivial axioms; while IR is not since it is violated by both the welfare methods proposed by Bernheim & Rangel (2009) and by Horan & Sprumont (2016). Finally, we say that a method is Relevant if it avoids "paradoxical" situations. We show how IR avoids two particular cases: (i) indisputable preferences are failed to be identified; (ii) the welfare relation becomes coarser and coarser when the number of observations increase.[26]

To show the Relevance of IR we consider a case in which the resulting preference order is indisputable and show that it can be inferred only by methods that satisfy such property. Suppose an individual evaluates the alternatives according to a utility function $u : X \to R_{++}$. At the act of choice, this utility is perturbed by an additive error component such that the choice depends on the random utility $U(x) = u(x) + \epsilon(x)$ where $\epsilon(x)$ is identically, independently and continuously

[24] For an analysis of non-neutral methods see Apesteguia & Ballester (2015).

[25] The necessity regards the normative principle. Our version of IR requires that only one observation is needed to break the indifference relation. The reader may want to define a weaker property where more than one observation is needed, still respecting the normative principle, but allowing for a more conservative welfare analysis.

[26] A binary relation is coarser than another one if it has a lower number of asymmetric parts.

distributed. The probability that x is chosen from a set $A \subseteq X$ will be $Pr[x = \mathsf{argmax}_{x \in A} U(x)]$. Furthermore, suppose that the dataset is restricted to multiple observations on a single set A; we denote this dataset dataset as $\mathcal{A}$. We show that given this particular restriction on the dataset, our three axioms can correctly identify the underlying deterministic utility u and consequently the correct welfare relation.

Proposition 1. *Given an i.i.d. RUM, a resulting collection of observations on a dataset $\mathcal{A}$ and a method that satisfies IR, NEU and CNN then $xR^{\mathcal{A}}y$ if and only if $u(x) \geqslant u(y)$.*

Proof. See Appendix A.1. □

A crucial part of the proof relies on the constraint posed on the domain $\mathcal{A} \subset \mathcal{D}$. This restriction could seem extremely severe. A more general result can be proved for a larger set of datasets at the cost of requiring the resulting binary relation to be transitive. Nonetheless, a weaker restriction has to be maintained. Particularly, a dataset $D \in \mathcal{D}$ is *homogeneous*, denoted as $\mathtt{hom}(D)$, if any $S \subseteq X$ with the same cardinality is observed the same number of times, which is assumed to be large.

Proposition 2. *Given an i.i.d. RUM, a resulting collection of observations over a dataset $\mathtt{hom}(D)$ and a method g that satisfies IR, NEU, CNN and Transitivity then $xR^{\mathtt{hom}(D)}y$ if and only if $u(x) \geqslant u(y)$.*

Proof. See Appendix A.2.[27] □

A brief comment on these results is needed to explain the role of the constraint on the dataset structure. The axioms required to identify the underlying utility function have to be satisfied on the restricted family of datasets and not in general. In Proposition 2 we require Transitivity which is generally not satisfied by **CRP**. Nonetheless, **CRP** satisfies Transitivity on the particular datasets we consider and therefore it can identify the underlying utility. Formally, we say **CRP** is not a transitive method[28] when it is defined as a map from $\mathcal{D}$ to $\mathcal{R}(X)$. However, it is transitive when it is defined as map from $\mathtt{hom}(\mathcal{D})$ to $\mathcal{R}(X)$.

2.3 Counting Procedures

So far, we have sustained the necessity of IR using counting procedures on particular domains. Now, we want to provide a general characterization of these procedures. In reality researchers deal with dataset that may have missing data or multiple observations. These features are embedded in our definition of dataset since the mapping $D \in \mathcal{D}$ may assign a zero value to all

[27] We prove this result for a collection of observations from an i.i.d. RUM; however the result can be extended to a larger family of stochastic models.

[28] There exists a dataset $D \in \mathcal{D}$ such that $R^D_{\mathbf{CRP}}$ is not transitive.

elements in a particular set (missing data) or assign different strictly positive values to more than one element (multiple observations). The following axioms are needed:

Axiom 4 (Independence **[IND]**)**.**

$$\forall \quad z \neq x, y \quad xR_f^D y \quad \Leftrightarrow \quad xR_f^{D+(z,A)} y$$

This axiom implies that the welfare relation between x and y does not depend on any observation where an element z is chosen. This is a strong axiom. In our experimental analysis we see that this axiom is rejected as necessary condition; namely there exist some methods that do not satisfy it and perform well.

Axiom 5 (Stability **[ST]**)**.**

$$\forall \ z \in X \quad xP_f^D y \quad \Rightarrow \quad \neg y P_f^{D+(z,A)} x$$

This axiom deals with the excessive sensitivity of the welfare method around the indifference. It states that one single observation cannot completely reverse the judgement. The stated version, limited to one observation, is the strongest possible given our structure. One may think of weaker versions that allow a reversal for observations that are considered of particular importance without changing the normative principle of this axiom. As for IND; our experimental analysis shows the non necessity of this axiom. It will be also interesting to note that the method proposed by Bernheim & Rangel (2009) trivially satisfies this axiom.

Axiom 6 (Strong Informational Responsiveness **[SIR]**)**.**

$$xI_f^D y \quad \Rightarrow \quad xP_f^{D+(x,A)} y$$

Axiom 7 (Connection **[CON]**)**.**

$$\forall \quad z \in X \quad \& \quad \forall \ A \not\supseteq \{x, y\} \quad xR_f^D y \quad \Leftrightarrow \quad xR_f^{D+(z,A)} y$$

These two axioms are important because they are the difference between the simple counting and the counting revealed preference procedure. Namely, **CC** satisfies SIR but not CON; while **CRP** satisfies IR and CON but not SIR. We are now ready to prove our two main results regarding counting procedures.

Theorem 1. *A welfare method satisfies ST, IND, SIR and NEU if and only if it is the simple counting method - [**CC**].*

Proof. See Appendix A.3. □

Few axiomatizations[29] of the simple counting have been provided but none of them deal with the complex dataset that we study in this paper. To see how the complexity of the dataset forces us to introduce ST, which is a novel axiom, consider a condition above mentioned called Positive Responsiveness. Formally, it is defined as SIR $\cap$ CNN. We show not only that this axiom is not sufficient together with IND and NEU but also that even adding transitivity [T] we cannot prove the statement without ST. The former case, perhaps less intuitive, is presented in Appendix B when we deal with the independence of the axioms. The latter is presented in the following example:

Example 1. *For all $x \in X$ and $D \in \mathcal{D}$:*

$$Q_x \geqslant Q_y \quad \Leftrightarrow \quad xR^D y$$

where $Q_x = \sum_{A \ni x} |A| \cdot D(x, A)$.

This welfare method gives more weight to sets with higher cardinality therefore violating ST. However, it is strictly monotonic in individual choices and since it maps into positive integers it satisfies transitivity as well.

In the following theorem, we constrain the counting procedure over the revealed preference relation. The reader may notice that IND is satisfied by this welfare method but implied by the other axioms, hence redundant.

Theorem 2. *A welfare method satisfies ST, IR, NEU and CON if and only if it is the counting revealed preference method - **[CRP]**.*

Proof. See Appendix A.4. □

2.4 Methods

We describe concisely the remaining methods that we test in our experiment. This list is comprehensive of all methods that, to be best of our knowledge, have been studied in the literature and can fit our abstract framework.

The methods are denoted as follows: **SEQ** is the *sequential method* - Horan & Sprumont (2016), **BR** is the *Bernheim and Rangel method* - Bernheim & Rangel (2009); **MS** is the *minimum swaps method* - Apesteguia & Ballester (2015), **EIG** is the *eigenvector centrality method*; **TC** is a variation of the *transitive core method* - Nishimura (2017); **OW** is the optimal weighted method.

[29] Rubinstein (1980) proposes an axiomatization for tournaments while van den Brink & Gilles (2003) for outdegree of digraphs.

2.4.1 Sequential

The sequential method can be effectively tested only if the dataset is constrained on $\mathcal{X}$ and it has one observation for each set.[30] It works recursively such that the best element is the one chosen from the universal set; the second best is the one chosen when the best alternative is removed; and so on.

Formally, let the dataset be constrained to $\mathcal{X}$ and only one observation is collected for each non-empty subset. We write $xP^D_{\mathbf{SEQ}}y$ for all $y \neq x$, if $D(x, X) = 1$; then $yP^D_{\mathbf{SEQ}}z$ for all $z \neq x, y$ if $D(y, X\backslash\{x\}) = 1$; again $zP_{\mathbf{SEQ}}w$ for all $w \neq x, y, z$ if $D(z, X\backslash\{x, y\}) = 1$ and so on.

2.4.2 Bernheim and Rangel

Bernheim & Rangel (2009) proposed the following method: x is (strictly) unambiguously better than y if y is never chosen when x is available. The method is acyclic when constrained on $\mathcal{X}$ and with no missing data.

Formally, $xP^D_{\mathbf{BR}}y$ if and only if for all $A \subseteq X$ such that $x, y \in A$, we have $D(y, A) = 0$.

2.4.3 Minimum swaps

The method has been proposed by Apesteguia & Ballester (2015). We denote it as **MS** and it is defined as the preference relation P that minimize a swaps index;[31] namely the number of alternatives that are ranked above the chosen one according to P. It may happen that more than one asymmetric binary relation P minimizes the above problem. In such case, we adopt the convention of taking the intersection among all the minimizers.

There is a strict connection between **CRP** and **MS** as it has been noted by Apesteguia & Ballester (2015).[32] We show that if $P^D_{\mathbf{CRP}}$ satisfies acyclicity then the transitive closure of $P^D_{\mathbf{CRP}}$ is equivalent to the asymmetric part of the minimum swaps relation $P^D_{\mathbf{MS}}$. This result is empirically important since it defines an equivalence between the asymmetric part of these methods for not-heavily irrational subjects; namely subjects that have an acyclic $P^D_{\mathbf{CRP}}$. In fact

[30] Horan & Sprumont (2016) suggest a way to extend the method over different datasets simply taking the intersection of all possible resulting orderings. Even though we use this methodology to prove that this method violates IR, we do not apply it empirically. The reader may note that this extension would not provide any additional and positive information to the empirical analysis of this method.

[31] Formally, the swaps index is defined as follows:

$$I_s(D, P) = \sum_{(x,A)\in O} |\{y \in A : yPx \;\&\; (x, A)\}|$$

[32] Apesteguia & Ballester (2015) introduced the following property: A collection of observations satisfies P-Monotonicity if xPy implies $C_{xy} > C_{yx}$. They then established the following result:

Theorem. *If a collection of observations satisfies P-Monotonicity, then P is the unique minimum swaps preference.*

we observe that almost all subjects are of this type: on the entire dataset we find only one subject with a cycle in time preferences and four in risk preferences.

Proposition 3. *If $P^D_{\boldsymbol{CRP}}$ is acyclic, then $xP^{*D}_{\boldsymbol{CRP}}y \Leftrightarrow xP^D_{\boldsymbol{MS}}y$; with P^* being transitive closure of P*

Proof. See Appendix A.5. □

2.4.4 Eigenvector centrality

This method uses the definition of centrality in networks in order to define an order of alternatives. Firstly, we construct the weighted revealed preference graph using C_{xy}. The eigenvector centrality of the nodes in the graph constitutes a complete and transitive ranking that measures the importance of each alternative.

2.4.5 Transitive core

This method has been recently proposed by Nishimura (2017). We introduce a variation of his original proposal which was in line with Bernheim & Rangel (2009). Instead, we found his approach on the **CRP** method. The transitive core method, denoted as **TC** is defined as follows:

$$xR^D_{\mathbf{TC}}y \quad \Leftrightarrow \quad \begin{cases} zR^D_{\mathbf{CRP}}x \Rightarrow zR^D_{\mathbf{CRP}}y \\ \\ yR^D_{\mathbf{CRP}}z \Rightarrow xR^D_{\mathbf{CRP}}z \end{cases} \quad \forall \quad z \in X$$

2.4.6 Optimal Weights

To define this method we divide the dataset in five parts, $i \in \Gamma$: binary sets [B], ternary sets [T], quaternary set [Q], sets with asymmetric dominance [AD], big sets [BIG]. For each part the revealed preference is collected creating, for each $x, y \in X$, a vector $\mathbf{C}_{xy} = (C^B_{xy}, C^T_{xy}, C^Q_{xy}, C^{AD}_{xy}, C^{BIG}_{xy})$. The weights vector is $\mathbf{w} = (w_B, w_T, w_Q, w_{AD}, w_{BIG})$. We define the method OW as follows:

$$xR^D_{\mathbf{OW}}y \text{ if and only if } OW_{xy} \geqslant OW_{yx}$$

where $OW_{xy} = \sum_{i \in \Gamma} w_i C^i_{xy}$.

Weights are calculated optimizing the sum of two measures: (1) expected identification of maximal element; (2) unique identification of the entire welfare relation. The former measures the expected number of subjects for whom the method can identify the reported best element; the latter measures the number of subjects for whom the method uniquely identify the entire reported welfare relation.[33]

[33]The optimality problem is performed using different objective functions in Subsection 4.5.

Given the definitions of identification procedures in Section 4.3, the optimization problem is as follows:

$$\max_{\mathbf{w}\in[-0.4,1]^5} \mathbf{EI} + \mathbf{WRI}$$

where

$$xR^D_{f_i}y \quad \Leftrightarrow \quad \mathbf{w}\cdot C_{xy_i}(\texttt{part}) \geqslant \mathbf{w}\cdot C_{yx_i}(\texttt{part})$$

2.4.7 Summary

Figure 1 summarizes the characteristics of the methods. Notice that, **OW** have missing properties. The reason is that since the optimal weights depend on both the dataset and the reported welfare relation we cannot say, a priori, if this method will satisfy some of the properties. If weights are negative then CNN is violated; if some of them are zero then both IR and SIR are violated. ST is almost always violated since the only case where it is satisfied is when **OW** is reduced to the **CRP** method. If this method satisfy IR then it would be evidence of the necessity of this property; we deal with this problem in Subsection 4.5.

It is important to notice that throughout the empirical analysis we substitute incompleteness with indifference. These process, that allows a consistent comparison across methods, can undermine the theoretical foundations of some of these methods. Particularly, **MS** and **TC** are affected; although differently. Both methods satisfy IR; however **MS** satisfies it even when indifferences are introduced; while **TC** does not. Therefore, we treat **MS** with indifferences and **TC** with incompleteness. Hence, **TC** satisfies both transitivity [T] and quasi-transitivity [QT]; while **MS** satisfies only QT.

It is not trivial to show that **TC** satisfies IR. The statement is denoted as Claim 1 and the proof can be found in Appendix A.6.

	NEU	CNN	IR	SIR	IND	ST	CON	QT	T
CRP	√	√	√	×	√	√	√	×	×
MS	√	√	√	×	×	×	×	√	×
TC	√	√	√	×	×	×	×	√	√
EIG	√	√	√	√	×	×	×	√	√
CC	√	√	√	√	√	√	×	√	√
SEQ	√	√	×	×	×	×	√	√	√
BR	√	√	×	×	√	√	√	×	×
OW	√	−	−	−	√	−	√	−	−

Figure 1: Summary of the properties of welfare methods.

3 Experimental design

The experiment follows a standard choice elicitation design, e.g. Manzini & Mariotti (2010), Barberá & Neme (2017). The complete instructions and screenshots are presented in the Online Appendix. Subjects received instructions both on screen and on paper such that they could consult them during the experiment.

The experiment is divided in three parts: (1) Choice elicitation part; (2) Questionnaire; (3) Raven Test. The choice elicitation part has 50 questions; half regarding choice among lotteries (Risk Preference Elicitation) and half regarding choice among delayed payment plans (Time Preference Elicitation); no question was repeated. At the beginning of each part subjects answered three trial questions in other to make them familiar with the experimental environment.

For both Time and Risk the alternatives were divided in two groups: four MAIN alternatives, that are presented in Table 1 and Table 2, and some "confounding" alternatives that are described in Online Appendix. Each individual solved all the 11 choice problems involving the MAIN alternatives. The other questions were set in order to obtain particular information about rationality: Monotonicity, Impatience,[34] Stochastic Dominance; and about possible behavioural effects: choice overload, compromise effect, attraction effect. The position of the alternatives were randomized. The subjects could face two orders of questions and also we inverted Time and Risk elicitation such that we had a total of four treatments.[35]

After the choice elicitation part subjects were asked, non-incentivized, to rank the four

[34]By Impatience we intend the violation of discounting models. The term "impatience" has been used by Fishburn & Rubinstein (1982) to denote Axiom A3.

[35]Given the high number of questions we apply a "structural randomization". Namely, we divide questions in groups by similarity and then we completely randomize with the constraints that similar questions could not appear clustered together.

Table 1: LIST OF MAIN DELAYED PAYMENT PLANS

ALTERNATIVES	MONTHS				
	0	**3**	**6**	**9**	**12**
One Shot (OS)	160	0	0	0	0
Decreasing (D)	110	50	25	0	0
Constant (K)	50	50	50	50	0
Increasing (I)	0	15	40	170	0

Table 2: LIST OF MAIN LOTTERIES

ALTERNATIVES	TOKEN		PROBABILITIES		EV
Degenerate (D)	50	0	1	0	50
Safe (S)	65	25	0.8	0.2	57
Fifty-Fifty (50)	90	25	0.5	0.5	57.5
Risky (R)	300	5	0.2	0.8	64

NOTES -- The amounts are described in Token. The exchange rate was fixed at 20:1 pounds for Delayed Payment Plans and 10:1 pounds for Lotteries.

MAIN alternatives. No indifferences were permitted, hence the reported welfare relation is always a linear order.[36] Subsequently, subjects filled a questionnaire containing questions about the comprehension of the experimental design and criteria of choice in both time and risk. The questionnaire is presented and analysed in Section 4 of the Online Appendix. Finally, two well-known test of cognitive abilities were presented: (i) Frederick Test - (Frederick, 2005); (ii) a selection of ten Raven matrices. Response times were collected for each question in the choice elicitation part and the cognitive abilities tests.[37]

The average reward was about 19 £ per subject and the experiment lasted on average 1:15 hours. The reward was measured in Token with an exchange rate of 1:10 for lotteries and 1:20 for delayed payment plans. Subjects received no feedback about their earnings during the experiment. At the end of the experiment computers randomly picked from chosen delayed payment plans and lotteries, this latter was played out, and in a last screen informed subjects of their earnings in each part.

All sessions were conducted at University of St. Andrews between June and September 2019. Subjects were recruited voluntarily among undergraduate and postgraduate students. Eleven sessions were run for a total of 145 subjects. No subject participated in more than one session. The earnings had been paid via bank account at the end of the experiment and in successive dates in the future as specified both by the instructions and by the experimenter. The experiment was completely anonymous and all subjects signed a consent form where they agreed in providing UK bank account number and sort code.

[36] A linear order is a complete, transitive and antisymmetric binary relation.

[37] Since this experiment is part of a larger project, the analysis of cognitive abilities, response times and structural axioms is treated in a compendium paper.

4 Results

4.1 CRP and BR

We begin showing the main result of the paper. Table 3 presents the identification power of **CRP** and **BR** as fraction of subjects for whom the methods can correctly identify either the reported best element or the entire welfare relation. As the reader may note in Figure 1, the only difference between these methods is that the latter does not satisfy IR. **CRP** performs significantly better along all dimensions both in time and risk preferences. Notably, **BR** is a lowest bound for the identification since when a violation is observed data are simply ignored. This means that the difference is performed on subjects that violate the Weak Axiom of Revealed Preference and therefore is not trivial.

Table 3: CRP & BR - IDENTIFICATION

	TIME			**RISK**		
METHODS	WRI	UI	EI	WRI	UI	EI
CRP	0.61	0.87	0.88	0.24	0.59	0.61
BR	0.42	0.59	0.74	0.06	0.14	0.43

NOTES -- **CRP** is the counting revealed preference method; **BR** denotes Bernheim & Rangel method. The numbers represent the fraction of subjects for whom the two welfare methods provide the following three identification: (1) "WRI" - Welfare Relation Identification and it refers to the unique identification of the entire reported welfare relation; (2) "UI" - Unique Identification of the reported best element; (3) "EI" - Expected Identification of the reported best element.

4.2 Premise: do individuals consistently reveal welfare?

Figure 2 presents the distribution of WARP violations in time, risk and random behaviour.[38] Two observations catch the eye: (i) subjects violate WARP less in time than in risk and the difference is statistically significant ($p \approx 0$); (ii) subjects do not behave randomly, again significantly ($p \approx 0$).

The difference is not based only on the presence of a higher number of rational individuals in time. If we restrict our test on those subjects that violate WARP at least once we find that the difference is still highly significant ($p \approx 0$). This suggests a fundamental difference in the behaviour of the agents in the two environments.

The suspicions are confirmed in Figure 3 where we show a scatter plot of the number of WARP violations. As the reader may notice the correlation is very low and driven mainly by a small fraction of consistent individuals. Given this preliminary evidence, we will treat Time and

[38] Given that the questions of time and risk were slightly different a random subjects may have in general different numbers of violations; however the difference is negligible. In order to provide a fair comparison we focus solely on the MAIN alternatives since they account for most of subjects' choices.

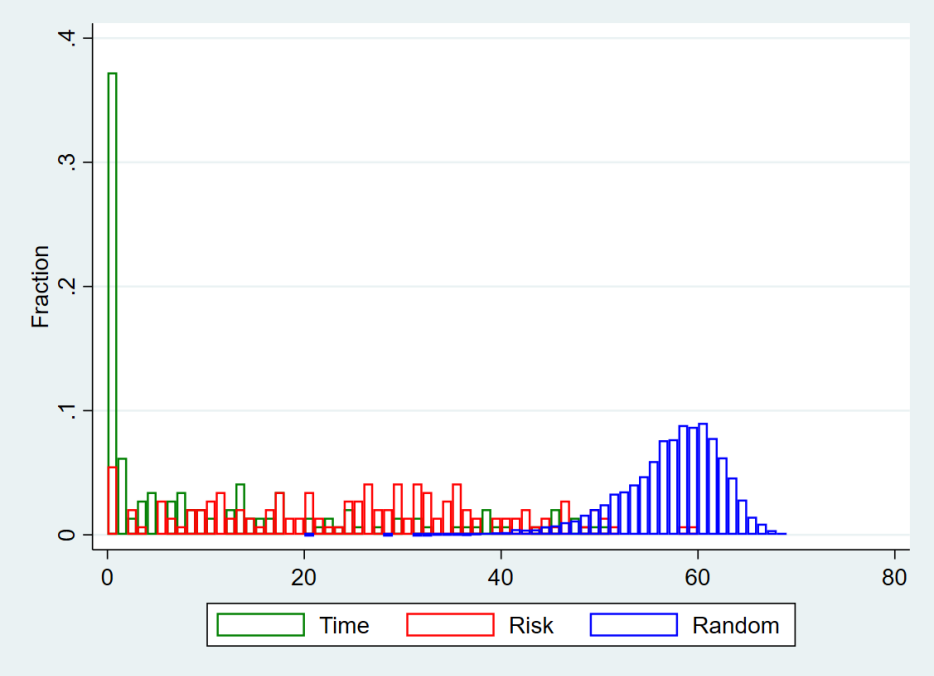

Figure 2: Distribution of the violations of WARP.

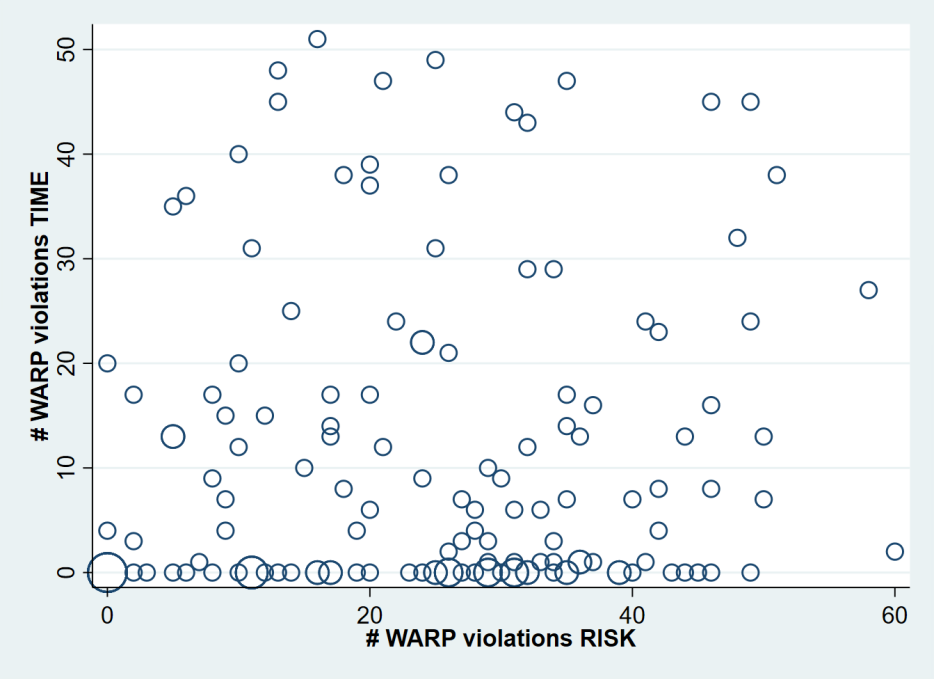

Figure 3: Scatter plot of the violations of WARP.

Risk consistency and preference elicitation analysis separately.

We begin investigating how violations are distributed in different parts of the dataset, that we call domains. In doing so we cannot rely simply on the number of violations of WARP since they depend on the number of questions and the alternatives. In other words we face the problem of: "... comparing the power of potentially different experimental designs. For a given

choice setting, some experimental designs may be more likely to reveal violations of GARP than others." - Andreoni et al. (2013). The problem can be rephrased as follows: suppose one subject makes 10 inconsistent choices among 40 binary choices while another subject makes 10 inconsistent choices among 30 ternary choices. How can we compare these subjects in terms of consistency?

A standard approach in evaluating consistency of individuals given different experiments is to compare them with random behaviour - see Becker (1962) and Bronars (1987). The literature has studied this problem starting from the notion of Selten measure (Selten, 1991) and has applied it to empirical studies such as in Beatty & Crawford (2011) and Echenique et al. (2011). We address the problem constructing a new index of consistency or "power index". We adopt the approach of perturbing a data generating process to create inconsistencies and compare the magnitude of the perturbation across domains.

As data generating process we build on the logit model as follows: let $A = \{x, y, z, w\}$ be the set of MAIN alternatives ordered by a linear order $\succ$ and u a utility function with $u(i) = u(j) + 1$ with $i, j \in A$ being consecutive elements in $\succ$. Note that, only differences in utility are important;[39] however the parameter identification is not invariant to positive affine transformations of u (not cardinal). The standard logit formula is the following:

$$p(x, A) = \frac{e^{u(x)}}{\sum_{y \in A} e^{u(y)}}$$

As in Train (2009)[40] we can modify the logit using a scale parameter λ connected to the variance of the unobserved error (a subject who chooses randomly behaves as if $\lambda = \infty$ but given our parameters for $\lambda \approx 5$ we substantially observe random behaviour); such that the formula becomes:

$$p(x, A) = \frac{e^{\frac{u(x)}{\lambda}}}{\sum_{y \in A} e^{\frac{u(y)}{\lambda}}}$$

The parameter λ can be also interpreted as the cost of acquiring information regarding the utility of the elements, e.g. Caplin & Dean (2015) and Fudenberg et al. (2015).

[39] Since in some part of the dataset the domain is not symmetric, namely some alternatives are more present than others. We adopt the convention of setting the utility difference of D and I (respectively S and R) equal to two. This is based on the fact the most of the subjects indicate in the ordinal ranking that these alternatives are divided by two positions; in particular either $OS \succ D \succ K \succ I$ or $I \succ K \succ D \succ OS$. We also ignore confounding alternatives since they account for a marginal part of the choice distribution in any sets where MAIN alternatives are also present.

[40] An example of maximum likelihood estimate of the paraemter λ can be found in McKelvey & Palfrey (1995). They show that in a game theoretical experimental (quantal response equilibria) setting subjects tend, with experience, to make less noisy choices.

We run a Monte Carlo simulation to estimate the parameter λ that match the average number of violations of WARP that the subjects make in the different part of the dataset. We only consider the MAIN alternatives since, as presented in Table 4 and 6, most of the violations, and choices, regard these alternatives.[41] Importantly this is not an estimation exercise (we do not believe that, when aggregated, subjects can be studied using a logit model). We provide an intuitive index that can be used for meaningful comparisons across domains. Given the strong assumptions made we also report the percentage of rational individuals and the standard deviation of our logit simulations such that the reader may have an idea of how close they are to the real data. We now present and comment the consistency analysis in Time and Risk.

4.2.1 Time

Table 4 shows the mean and standard deviation of the number of WARP violations within different parts of the dataset; as well as the percentage of rational individuals, namely those with zero violations. Since these numbers are not comparable we look at the logit index. It shows that questions with asymmetric dominance effect present a relatively higher number of violations. The difference between BIG sets and MAIN sets is instead very small. To give an idea of how measures of rationality can be misread, the reader may note that the percentage of rational subjects in AD sets is biased by the fact that only four questions have this characteristic[42] making highly probable for mildly irrational subjects to report zero violations.

Table 4: WARP VIOLATIONS - TIME

	BIG	AD	MAIN	ALL**	ALL
Mean	1.4897	0.8138	1.9586	10.0621	11.2621
Std	1.9189	1.4577	2.9009	14.2099	14.5263
Rational	54%	75%	59%	48%	37%
Logit - λ	0.555	0.787	0.515	0.569	-
Logit - std	1.6406	1.3738	2.0355	7.768	-
Logit - Rational	48%	74%	40%	16%	-

NOTES -- The mean of WARP violations is reported for different parts of the dataset: "BIG" denotes sets with more than 8 elements; "AD" denotes sets with potential asymmetric dominance effect; "MAIN" denotes the 11 non-empty subsets of the four main alternatives; "ALL" denotes the entire dataset. ALL** refers to WARP violations in the entire dataset that regard only the four main alternatives. We also report the following statistics: the information parameter of a logit model that match the data mean, the standard deviation and percentage of rational subjects in the resulting distribution.

[41] This result is evidenced by the small difference between the violation in ALL** and ALL datasets. This assumption is conservative; in fact in AD or BIG sets the identification of the parameter λ is lower than it would be.

[42] Since we ignore dominated alternatives, the simulation uses a dataset made of four binary sets of the type $\{D, I\}$. Considering dominated alternatives would force even more ad hoc evaluations of the utility functions.

Table 5: WARP VIOLATIONS - TIME

	MAIN/BIG	MAIN/AD	BIG/AD
Mean	3.2897	1.3724	1.9172
Std	4.6682	2.5568	2.8052
Rational	56%	73%	56%
Logit - λ	0.515	1.062	1.124
Logit - std	2.8921	2.0696	2.508
Logit - Rational	34%	66%	59%

NOTES -- The mean of WARP violations is reported between difference domains: "MAIN/BIG" denotes violations observed between MAIN and BIG sets; "MAIN/AD" denotes violations between MAIN and AD sets; "BIG/AD" denotes violations between BIG and AD sets. These numbers are calculated, for instance, taking the total number of violations on MAIN and BIG sets and subtracting the violations within the two domains.

Two observations are worth noting. (1) Higher is the number of sets and worse is the logit approximation to the data. For instance, on the entire dataset we should observe 16% of rational subjects while we observe 48% and the standard deviation is also significantly higher. (2) The coefficient of variation is everywhere above one. This evidence suggests that there are at least two different groups of subjects: one rational and the other irrational; importantly this latter has been shown to behave not randomly.

Table 5 shows the number of violations of WARP between different domains; for instance when x is chosen over y in one of the MAIN sets and y over x in a one of the BIG sets. The results show that not only the level of rationality is similar between MAIN and BIG sets but also the types of violations are similar. In fact, AD sets present a different behaviour from both the other domains; namely to match the number of WARP violations between AD sets and the other domains we would require a level of perturbation higher than all levels within the domains. Furthermore, Table 5 confirms the presence of at least two groups of individuals since the standard deviation of the logit simulations is everywhere below the standard deviation in the data.

4.2.2 Risk

Table 6 reports the results regarding WARP violations within domains in risk preferences. Firstly, the number of violations is everywhere higher than in time preferences across all the domains and everywhere significantly ($p \approx 0$). In this case, the comparison between time and risk environment is meaningful given the approximate symmetry of the datasets. This evidence suggests that the difference in behaviour between the two environments is not due to particular incidence of behavioural effects. The difference in the shape of the distribution expressed in Figure 2 is confirmed by the coefficients of variation. In Time they are everywhere bigger than

one, confirming that the left skewed shape is a common property across domains, while in Risk they are almost everywhere smaller than one, confirming the uniform shape of the distribution of WARP violations. Surprisingly, Table 6 shows that in BIG sets subjects have a more rational behaviour compared to both AD and MAIN sets.[43]

Table 6: WARP VIOLATIONS - RISK

	BIG	AD	MAIN	ALL**	ALL
Mean	4.6690	1.2621	4.9862	21.7172	24.6552
Std	2.8700	1.4955	3.4237	13.6314	14.3170
Rational	15%	54%	14%	8%	6%
Logit - λ	0.756	1.003	1.009	0.774	-
Logit - std	2.9899	1.5672	2.6606	9.6465	-
Logit - Rational	22%	60%	7%	2%	-

NOTES -- See Table 4.

Table 7: WARP VIOLATIONS - RISK

	MAIN/BIG	MAIN/AD	BIG/AD
Mean	9.0276	2.3241	2.3862
Std	5.9224	2.7267	2.5888
Rational	14%	44%	40%
Logit - λ	0.688	1.125	1.581
Logit - std	4.7995	2.2564	2.3141
Logit - Rational	9%	37%	32%

NOTES -- See Table 5.

Thirdly, we confirm that when the number of sets increase data shows a percentage of rational subjects higher than the logit simulation as well as a much higher standard deviation. Finally, Table 7 shows a higher similarity in the behaviour of subjects in MAIN and BIG sets compared to both MAIN/AD and BIG/AD sets. It is particularly interesting to notice the extremely high logit index associated with violations between BIG and AD sets. Speculations would lead us to conjecture that choice overload and asymmetric dominance, although both in the family of behavioural effects, have very different implications on the consistency of behaviour in choice among lotteries.

[43]This evidence may be related with attention models such Masatlioglu et al. (2012), Manzini & Mariotti (2014a), Lleras et al. (2017) and Cattaneo et al. (2018), and could confirm previous experiments such as Iyengar & Kamenica (2010). On the contrary models that assume more uniform stochastic choice in BIG sets such as Fudenberg et al. (2015) and Frick (2016) seem to be not backed by the data.

4.3 Identification of reported welfare

This subsection contains the main results of the paper. We measure the power of identification of different welfare methods in both time and risk using ALL dataset, MAIN sets and BINARY sets. This latter is considered as a benchmark to understand how much information can be extracted outside a dataset that does not present any potential behavioural effect. Two results emerge in both Time and Risk: (1) methods that satisfy IR performs significantly better than **BR**; (2) the identification power of methods that satisfy IR improves when more data are collected. This result, as expected, is reversed in **BR**.

Our identification exercise is threefold. Firstly, we uniquely identify the reported best element. Secondly, since **BR** is a conservative approach, it is reasonable to imagine that this method performs better in a set identification exercise; namely when the reported best element is in the set of maximal elements. We assume that a risk-neutral policy maker has to pick from the set of maximal elements endowed with a uniform distribution. Given this assumption, we perform an expected identification exercise. Finally, we uniquely identify the entire reported welfare relation.

Let N be the set of subjects and $f_i(D)$ be the preference elicited by the welfare method f given the choices of subject i over the dataset D. The reported welfare relation by subject i is denoted as $\mathsf{REP}_i(\succ)$. The proportion of correctly identified subjects given the three approaches is as follows:

- Unique Identification [**UI**]:

$$\frac{\#\{i \in N : \max[\mathsf{REP}_i(\succ)] = \max[f_i(D)]\}}{\#N}$$

- Expected Identification [**EI**]:

$$\frac{\sum_{i \in N : \max[\mathsf{REP}_i(\succ)] \in \max[f_i(D)]} \frac{1}{\#\{\max[f_i(D)]\}}}{\#N}$$

- Welfare Relation Identification [**WRI**]:

$$\frac{\#\{i \in N : \mathsf{REP}_i(\succ) = f_i(D)\}}{\#N}$$

Note that, the reported welfare relation is necessarily asymmetric. Hence, methods that map into linear orders such as **SEQ** or **EIG** are theoretically favoured in the identification of the entire welfare relation. To solve this issue we also investigate how close methods are to identify reported welfare relation even when these are not perfectly identified. The similarity of solutions is measures using the sum over all subjects of: (1) the cardinality of the symmetric

difference between the resulting binary relations and the reported order;[44] (2) the number of times the asymmetric part of the reported order is reversed. Using both measures is crucial. The symmetric difference considers equally the symmetric and asymmetric part of the binary relation, hence punishing coarse methods such as **BR**. The "reverse asymmetry" measure allows us to disentangle those differences that are in principle worse; namely when a subject reports **x** better than **y** but the method ranks **y** better than **x**. This measure punishes particularly methods that map in linear orders such as **EIG** and **SEQ**; while the conservative nature of **BR** creates a lowest bound. This analysis, together with the three identification exercises, provide a comprehensive picture of the identification power of each method.

4.3.1 Time

Table 8 shows that methods that satisfy IR perform significantly better than **BR** both uniquely (30%) and in expectation (15%). It is crucial to notice that **BR** is a lowest bound in the identification exercise since it identifies only those subjects that rationally reveal their best element. Therefore, the 30% gap is not trivial because it is performed on irrational individuals.

Table 8: UNIQUE & EXPECTED IDENTIFICATION - TIME

		UNIQUE			EXPECTED		
METHODS		**ALL**	**MAIN**	**BINARY**	**ALL**	**MAIN**	**BINARY**
IR & CRP	**CRP**	0.87	0.81	0.77	0.88	0.84	0.77
	MS	0.87	0.81	0.79	0.88	0.85	0.80
	EIG	0.87	0.83	0.81	0.87	0.83	0.81
	TC	0.88	0.81	0.77	0.88	0.83	0.77
IR	**CC**	0.81	0.83	0.77	0.84	0.86	0.81
No-IR	**SEQ**	-	0.83	-	-	0.83	-
	BR	0.59	0.67	0.77	0.74	0.79	0.77
	OW	0.89	-	-	0.89	-	-

NOTES -- On the left we show the portion of subjects for whom each method uniquely identify the reported best element. On the right, the expected portion of subjects for whom each method identify the reported best element. The measure is expected because for some subjects methods may set identify the best element; in these cases we assume to pick uniformly from the set of identified elements.

The power of identification for methods that satisfy IR is increasing in the number of sets in the dataset which suggests that individuals reveal information about welfare along all the dataset. Only exception is **CC**. We interpret as evidence in favour of the importance of standard revealed preference as foundation for welfare methods.

[44] The symmetric difference $\triangle$ between two binary relations R_1, R_2 is defined as follows: $R_1 \triangle R_2 = (R_1 \backslash R_2) \cup (R_2 \backslash R_1)$. For instance, let $R_1 = \{(x,y),(y,x),(y,z),(x,z)\}$ and $R_2 = \{(x,y),(y,z),(z,y),(x,z)\}$ we have $R_1 \triangle R_2 = |\{(y,x),(z,y)\}| = 2$.

Finally, **SEQ** performs particularly well; the difference is only 4-6%. The reason is that the best element of **SEQ** is the one chosen from the set with all the four main alternatives. It turns out this choice is a good predictor of the reported best element, although the two elicitations are not equivalent.

Table 9 shows the identification of the entire welfare relation. We present it together with symmetric difference and reverse asymmetry measures.

Table 9: IDEN. WELFARE RELATION, SD & RA - TIME

		ENTIRE IDEN.			SD & RA					
		ALL	**MAIN**	**BINARY**	**ALL**		**MAIN**		**BINARY**	
	METHODS	-	-	-	SD	RA	SD	RA	SD	RA
IR & CRP	**CRP**	0.61	0.57	0.59	180	78	191	73	220	110
IR & CRP	**MS**	0.62	0.59	0.61	182	82	188	76	218	88
IR & CRP	**EIG**	0.54	0.60	0.61	222	111	208	104	218	109
IR & CRP	**TC**	0.61	0.58	0.59	180	73	188	68	234	71
IR	**CC**	0.54	0.58	0.59	214	91	186	74	218	78
No-IR	**SEQ**	-	0.60	-	-	-	194	97	-	-
No-IR	**BR**	0.42	0.50	0.59	264	45	226	54	220	110
	OW	0.66	-	-	170	85	-	-	-	-

NOTES -- On the left we show the portion of subjects for whom each method uniquely identify the entire reported welfare relation. On the right, "SD" and "RA" denote respectively symmetric difference and reverse asymmetry.

We confirm that methods that satisfy IR perform better than **BR** by 10-15%. The performances of **SEQ** and **EIG** are positively biased by the feature that they do not allow indifferences. In fact, observing the measure of RA we see that they significantly reverse more asymmetric parts than the other methods. Since **BR** constitutes a lower bound in RA, and setting it to zero, we can say that for the all dataset they perform worse than **MS** by respectively 30% and 65%.[45]

The monotonicity of the identification power in the size of the dataset is not straightforward. However, if we observe the SD of methods that satisfy IR we notice that it is decreasing for any method apart from **EIG** and **CC**. This latter result was expected; while the poor performance of **EIG** is due to both the absence of indifference and the excessive weight posed by the method on observations from big sets.

4.3.2 Risk

Table 10 shows that methods satisfy IR perform significantly better than **BR** both uniquely (50%) and in expectation (20%). We also confirm that the power of identification is generally

[45] We consider **MS** since it the method that in this case maximizes the identification of the entire reported welfare relation. We do not consider **OW** since it was, at least partly, designed for this purpose.

(note that **CC** is still an exception) increasing in the size of the dataset.

The choice from the set of main alternatives is again a good predictor of the reported best element since the loss of **SEQ** is only 4-8%.

Table 10: UNIQUE & EXPECTED IDENTIFICATION - RISK

		UNIQUE			EXPECTED		
METHODS		**ALL**	**MAIN**	**BINARY**	**ALL**	**MAIN**	**BINARY**
IR & CRP	**CRP**	0.59	0.52	0.42	0.61	0.59	0.42
	MS	0.59	0.52	0.46	0.61	0.60	0.50
	EIG	0.61	0.61	0.51	0.61	0.61	0.51
	TC	0.61	0.51	0.42	0.62	0.55	0.42
IR	**CC**	0.55	0.56	0.42	0.57	0.61	0.50
No-IR	**SEQ**	-	0.55	-	-	0.55	-
	BR	0.14	0.25	0.42	0.43	0.49	0.42
	OW	0.63	-	-	0.63	-	-

NOTES -- See Table 8.

Table 11 again shows that methods that satisfy IR outperform **BR** in the entire identification exercise by 15-20%. We also confirm that **SEQ** and **EIG** performances are only apparently good; in fact when controlled for RA measure, and normalizing for the RA measure of **BR**, we see that they perform worse than **MS** by respectively 25% and 17%.

The monotonicity in the identification power is confirmed both looking at SD and at the identification process. It is interesting to notice that contrarily to time, the identification of **EIG** is increasing in the size of the dataset. This suggests either that big sets are important in identifying the entire welfare relation, or that binary sets are not important, or both. We investigate and confirm this hypothesis in the last subsection.

Table 11: IDEN. WELFARE RELATION, SD & RA - RISK

		ENTIRE IDEN.			SD & RA					
		ALL	**MAIN**	**BINARY**	**ALL**		**MAIN**		**BINARY**	
METHODS		-	-	-	SD	RA	SD	RA	SD	RA
IR & CRP	**CRP**	0.24	0.19	0.20	436	186	455	168	556	278
	MS	0.24	0.20	0.21	440	190	452	179	569	241
	EIG	0.30	0.27	0.23	446	223	448	224	576	288
	TC	0.24	0.19	0.20	434	182	446	157	570	184
IR	**CC**	0.21	0.19	0.20	453	200	452	185	569	218
No-IR	**SEQ**	-	0.25	-	-	-	478	239	-	-
	BR	0.06	0.10	0.20	592	86	545	115	556	278
	OW	0.32	-	-	421	210	-	-	-	-

NOTES -- See Table 9.

4.4 Completeness of the methods

In this section, we present the results related with the completeness of the methods. We borrow the term "completeness" from Fudenberg et al. (2019). The authors use machine learning to measure the amount of variation in the data that a theory can capture. The definition of completeness aims to answer the following question: "How close is the performance of a given theory to the best performance that is achievable in the domain?" Fudenberg et al. (2019). In our framework, we define completeness, denoted as $\texttt{Com}(f)$ for some welfare method f, as:

$$\texttt{Com}(f) = \frac{\varepsilon(f_L) - \varepsilon(f)}{\varepsilon(f_L) - \varepsilon(f_U)}$$

where $\varepsilon(f_L)$ is the proportion of non-identified subjects by the method that defines a lower bound on the domain; $\varepsilon(f_U)$ is the best achievable residual proportion and $\varepsilon(f)$ is the residual proportion of the model under study. In our framework, we set $f_L =$ **BR** and $f_U =$ **OW**. Table 12 shows the completeness of the methods using ALL sets across different types of identification procedures in both Time and Risk.

Table 12: COMPLETENESS OF THE METHODS

	TIME			RISK		
METHODS	UNI.	EXP.	ENT.	UNI.	EXP.	ENT.
CRP	0.93	0.93	0.79	0.92	0.90	0.69
MS	0.93	0.93	0.83	0.92	0.90	0.69
EIG	0.93	0.86	0.50	0.96	0.90	0.92
TC	0.95	0.93	0.79	0.96	0.95	0.69
CC	0.74	0.66	0.50	0.84	0.70	0.58
SEQ	0.81	0.59	0.75	0.84	0.60	0.73
BR	0.00	0.00	0.00	0.00	0.00	0.00
OW	1.00	1.00	1.00	1.00	1.00	1.00

NOTES -- This table reports the completeness of all methods in cases of unique (UNI.), expected (EXP.) and entire (ENT.) identification procedures.

Since **BR** and **OW** are respectively lower and upper bound for our identification analysis they take respectively value zero and one. Methods that satisfy IR and are based on the revealed preference approach have generally higher completeness than other methods. Note that, even though we do not report completeness for the measures of symmetric difference and reverse asymmetry in the entire identification approach, that favours **SEQ** over other methods, there always exists at least a method among those that satisfy IR and are based on revealed preference that is more complete than **SEQ**.

4.5 Informational Responsiveness & Optimal Weights

In Section 2 we propose IR as necessary condition for welfare methods. We can exploit one implication of IR to directly test the axiom. Note that, in the family of weighted sums, if revealed preferences receive strictly positive weights[46] in any part of the dataset then IR is satisfied. Hence, our construction of **OW** allows us to test whether IR binds in an optimal identification problem.

4.5.1 Time

Table 13 shows the intervals of weights that guarantee optimality for different objective functions. We generalize our previous analysis where the convention was to optimize the sum of expected identification of the reported best element and unique identification of the entire welfare relation. Since weights are often not unique, we report the minimum and maximum weights for which there exists a system of weights that solve the optimization problem. This does not imply

[46]This implication is immediate. See Meyer & Mongin (1995) for a comprehensive study of affine aggregation.

that any vector of weights that is in the cartesian product of the intervals guarantees optimal identification. In the table, for completeness of information, we split the MAIN sets in three parts: Binary sets, Ternary sets and Quaternary set.

Table 13: OPTIMAL WEIGHTS - TIME

	TIME				
IDENTIFICATIONS	**BIN**	**TER**	**QUA**	**BIG**	**AD**
UI	[0.6,1]	[0.2,1]	[0.1,0.2]	[0.6,1]	[-0.2,1]
EI	[0.6,1]	[0.2,1]	[0.1,0.2]	[0.6,1]	[-0.2,1]
WRI	[0.2,0.9]	[0.3,1]	[0.3,1]	[0.4,1]	[-0.2,-0.1]
SD	[0.5,0.8]	[0.6,1]	[0.4,0.8]	[0.4,0.7]	-0.2
SD & RA	0.6	0.6	0.6	0.6	-0.2
EI & WRI	0.9	1	0.4	0.8	-0.2

NOTES -- The table contains intervals of weights that optimize the identification of different objectives. "UI" and "EI" denote respectively unique and expected identification of the best element; "WRI" denotes entire welfare relation identification; "SD" and "RA" denote respectively minimization of the sum of symmetric difference and [two times] reverse asymmetry against the reported welfare relation; "EI & WRI" denotes the sum of EI and WRI. This latter is the one used along the paper to define OW.

We observe that strictly positive weights are associated to any part of the dataset apart from AD sets. This latter is found to be irrelevant in the identification of the reported best element (weights can be negative, zero or positive), while they have negative weights when we identify the entire welfare relation. This result is somewhat surprising since it shows that subjects wrongly reveal their welfare in this part of the dataset. Nonetheless, it confirms the findings of Section 4.2, where we show that subjects are not only more irrational in these sets (Table 4); but also they have different behaviour (Table 5) if compared to MAIN and BIG sets.

We also find that binary sets are particularly important along all the possible objective functions. This explains both the relatively good performance of methods on these sets (Table 8) and the fact that the identification power of **EIG** decreases in the size of the sets as observed in Table 9. This is due to the high weight put to bigger sets by the **EIG** method.

4.5.2 Risk

Table 14 shows that IR binds everywhere since strictly positive weights are attached to any domain. There are two exceptions. Firstly, AD sets are irrelevant when we focus only on the reported best element. This confirms the data in Table 7, where we show a different behaviour between AD sets and the rest of the dataset.

Table 14: OPTIMAL WEIGHTS - RISK

	RISK				
IDENTIFICATIONS	**BIN**	**TER**	**QUA**	**BIG**	**AD**
UI	[-0.2,0]	[0.4,0.7]	[0.7,1]	[0.5,0.9]	[-0.2,0.9]
EI	-0.1	[0.3,0.8]	[0.5,1]	[0.4,0.9]	[-0.2,0.9]
WRI	[0.2,0.7]	[0.4,1]	[0.8,1]	[0.3,0.8]	[0.3,1]
SD	0.4	0.4	1	0.3	0.4
SD & RA	0.5	[0.4,0.5]	[0.8,1]	[0.4,0.5]	[0.4,0.5]
EI & WRI	[0.1,0.3]	[0.4,0.5]	[0.8,1]	[0.4,0.6]	[0.3,0.6]

NOTES -- See Table 13.

Secondly, when we focus only on the identification of the reported best element we observe that binary sets receive weakly negative weights. These weights are also strictly positive but close to zero in the other exercises. This again confirms the findings of previous sections. In fact, in Table 10 we find that methods perform poorly on binary sets. We also found (Table 11) that the **EIG** method has an increasing identification power in the size of the sets. Finally in Table 11 we find that, throughout all methods, the differential of both symmetric difference and reverse asymmetry between binary sets and MAIN and ALL dataset is positive and significant.

The low importance of binary sets is striking. Especially, if we compare the weights associated with BIG sets where supposedly we should observe choice overload effect. This seems to suggest that, in risk, the irrational behaviour in MAIN sets is mostly driven by binary sets.[47]

5 Conclusion

In this paper, we axiomatically analyse welfare analysis. We propose normatively appealing properties and show that they have important empirical implications. Particularly, we propose a property called Informational Responsiveness. We show that it is a necessary condition to avoid paradoxical welfare conclusions and to satisfy the principle that more data should lead to finer conclusions. As a novelty we characterize the counting revea3aled preference procedure on datasets with possibly multiple observations and missing data. We argue that Informational Responsiveness together with a revealed preference approach are necessary conditions for an effective welfare analysis.

In the second part of the paper, using a novel experimental design, we test our hypothesis both in its premise and its conclusion. Firstly, we show that individuals repeatedly violate the Weak Axiom of Revealed Preference both in time and risk preferences. We develop a new index of rationality and show that inconsistency is a general phenomena, namely it is common to

[47]This evidence suggest further research on attentions in choice among gambles and it is in line with stochastic models such as Manzini & Mariotti (2014a) and Cattaneo et al. (2018).

sets with cardinality and with and without behavioural effects. Secondly, we find that welfare methods that satisfy Informational Responsiveness and are based on a revealed preference approach perform significantly better in identifying both the best reported element and the entire reported welfare relation. The results are strong in both time and risk preferences and in any part of the dataset. We show that these welfare methods are more complete theories in the sense of Fudenberg et al. (2019). Finally, using an optimal weighting algorithm we directly test Informational Responsiveness. We show that subjects reveal welfare in almost all part of the dataset and therefore that welfare analysis is most effective when all data are used but they are differently weighted according to the capacity of revealing welfare.

References

Afriat, S. N. (1973). On a system of inequalities in demand analysis: an extension of the classical method. *International Economic Review*, *14*(2), 460–472.

Agranov, M., & Ortoleva, P. (2017). Stochastic choice and preferences for randomization. *Journal of Political Economy*, *125*(1), 40–68.

Andreoni, J., Gillen, B. J., & Harbaugh, W. T. (2013). The power of revealed preference tests: ex-post evaluation of experimental design. *Working paper*.

Apesteguia, J., & Ballester, M. A. (2015). A measure of rationality and welfare. *Journal of Political Economy*, *6*(123), 1278–1310.

Arrow, K. J. (1959). Rational choice functions and orderings. *Economica*, *26*(102), 121–127.

Barberá, S., & Neme, A. (2017). Ordinal relative satisficing behavior. *Working paper*.

Beatty, T. K., & Crawford, I. A. (2011). How demanding is the revealed preference approach to demand? *American Economic Review*, *101*(6), 2782–2795.

Becker, G. S. (1962). Irrational behaviour and economic theory. *Journal of Political Economy*, (70), 1–13.

Bernheim, B. D., & Rangel, A. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *The Quartely Journal of Economics*, *124*(1), 51–104.

Bronars, S. G. (1987). The power of nonparametric tests of preference maximization. *Econometrica*, *55*(3), 693–698.

Caplin, A., & Dean, M. (2015). Revealed preference, rational inattention and costly information acquisition. *American Economic Review*, *105*(7), 2183–2203.

Cattaneo, M. D., Ma, X., Masatlioglu, Y., & Suleymanov, E. (2018). A random attention model. *Working paper*.

Cavagnaro, D. R., & Davis-Stober, C. P. (2014). Transitive in our preferences, but transitive in different ways: an analysis of choice variability. *Decision*, *1*(2), 102–122.

Chabris, C. C., Laibson, D., Morris, C. L., Schuldt, J. P., & Taubinsky, D. (2009). The allocation of time in decision making. *Journal of the European Economic Association*, (7), 628–637.

Danan, E., & Ziegelmeyer, A. (2006). Are preferences complete? an experimental measurement of indecisiveness under risk.

Echenique, F., Lee, S., & Shum, M. (2011). The money pump as a measure of revealed preference violations. *Journal of Political Economy*, *119*(6), 1201–1223.

Echenique, F., Saito, K., & Tserenjigmid, G. (2018). The perception-adjusted luce model. *Mathematical Social Sciences*, *93*, 67–76.

Famulari, M. (1995). A household-based, nonparamtetric test of demand theory. *The Review of Economics and Statistics*, (77), 372–83.

Fishburn, P. C., & Rubinstein, A. (1982). Time preference. *International economic review*, *23*(3), 677–694.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Frick, M. (2016). Monotone threshold representations. *Theoretical Economics*, (11), 757–772.

Fudenberg, D., Iijima, R., & Strzalecky, T. (2015). Stochastic choice and revealed perturbed utility. *Econometrica*, *83*(6), 2371–2409.

Fudenberg, D., Kleinberg, J., Liang, A., & Mullainathan, S. (2019). Measuring the completeness of theories. *PIER Working Paper No. 18-010*.

Goodin, R. E., & List, C. (2006). Special majorities rationalized. *British Journal of Political Science*, (36), 213–241.

Green, J., & Hojman, D. (2007). Choice, rationality and welfare measurement. *unpublished*.

Harbarugh, W. T., Krause, K., & Berry, T. R. (2001). Garp for kids: on the development of rational choice behavior. *American Economic Review*, *91*(5), 1539–1545.

Haynes, G. A. (2009). Testing the boundaries of the choice overload phenomenon: the effect of number of options and time pressure on decision difficulty and satisfaction. *Psychology & Marketing*, *26*(3), 204–212.

Hey, J. (2001). Does repetition improve consistency? *Experimental economics*, *4*(1), 5–54.

Hey, J., & Carbone, E. (1995). Stochastic choice with deterministic preferences: an experimental investigation. *Economics Letters*, *47*(2), 161–167.

Horan, S., & Sprumont, Y. (2016). Welfare criteria from choice: An axiomatic analysis. *Games and Economic Behavior*, (99), 56–70.

Houtman, M., & Maks, J. A. H. (1985). Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve Methoden*, (19), 89–104.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: violations of regularity and the similarity hypothesis. *The Journal of Consumer Research*, *9*(1), 90–98.

Iyengar, S. S., & Kamenica, E. (2010). Choice proliferation, simplicity seeking, and asset allocation. *Journal of Public Economics*, (94), 530–539.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: can one desire too much of a good thing? *Journal of personality and social psychology*, *79*(6), 995–1006.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). Foundations of measurement. *Volume I*.

Lleras, J. S., Masatlioglu, Y., Nakajima, D., & Ozbay, E. Y. (2017). When more is less: limited consideration. *Journal of Economic Theory*, (170), 70–85.

Manzini, P., & Mariotti, M. (2010). Revealed preference and boundedly rational choice procedures: an experiment. *Unpublished*.

Manzini, P., & Mariotti, M. (2012). Categorize then choose: boundedly rational choice and welfare. *Journal of the European Economic Association*, (10), 1141–1165.

Manzini, P., & Mariotti, M. (2014a). Stochastic choce and consideration sets. *Econometrica*, *82*, 1153–1176.

Manzini, P., & Mariotti, M. (2014b). Welfare economics and bounded rationality: the case for model-based approaches. *Journal of Economic Methodology*, (12), 343–360.

Marschak, J., & Block, H. (1960). Random orderings and stochastic theories of responses. *Contributions to Probability and Statistics*, (Stanford University Press).

Masatlioglu, Y., Nakajima, D., & Ozbay, E. Y. (2012). Revealed attention. *American Economic Review*, *102*(5), 2183–2205.

May, K. O. (1952). A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica*, *20*(4), 680–684.

McKelvey, R. D., & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and economic behavior*, *10*(1), 6–38.

Meyer, B. D., & Mongin, P. (1995). A note on affine aggregation. *Economics Letters*, *47*(2), 177–183.

Natenzon, P. (2019). Random choice and learning. *Journal of Political Economy*, *127*(1), 419–457.

Nishimura, H. (2017). The transitive core: inference of welfare from nontransitive preference relation. *mimeo*.

Rubinstein, A. (1980). Ranking the participants in a tournament. *SIAM Journal on Applied Mathematics*, *38*(1), 108–111.

Rubinstein, A., & Salant, Y. (2012). Eliciting welfare preferences from behavioural data sets. *The Review of Economic Studies*, (79), 375–387.

Salant, Y., & Rubinstein, A. (2008). (a,f): Choice with frames. *Review of Economic Studies*, *75*, 1287–1296.

Selten, R. (1991). Properties of a measure of predictive success. *Mathematical Social Sciences*, *21*(2), 153–167.

Sen, A. (1971). Choice functions and revealed preference. *The Review of Economic Studies*, *38*(3), 307–317.

Sippel, R. (1997). An experiment on the pure theory of consumer's behaviour. *Economic Journal*, *107*(444), 1431–1444.

Sopher, B., & Narramore, M. J. (2000). Stochastic choice and consistency in decision making under risk: an experimental study. *Theory and Decision*, *48*(4), 323–350.

Swofford, J. L., & Whitney, G. A. (1987). Nonparametric test of utility maximization and weak separability for consumption, leisure and money. *The Review of Economics and Statistics*, (69), 458–64.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, (76), 31–48.

Tversky, A., & Russo, J. E. (1969). Substitutability and similarity in binary choices. *Journal of Mathematical Psychology*, *6*(1), 1–12.

van den Brink, R., & Gilles, R. P. (2003). Ranking by outdegree for directed graphs. *Discrete Mathematics*, (271), 261–270.

A Proofs

A.1 Proposition 1

In the following proofs we omit the subscript f to ease the reading. By law of large numbers, we have $u(x) \geqslant u(y)$ if and only if $C_x \geqslant C_y$. We have to prove $xR^{\mathcal{A}}y$ if and only if $C_x \geqslant C_y$. The only if part is trivial since the counting choice method satisfies all three axioms.

Take two elements $x, y \in A$ and divide the dataset in three disjoint parts: C_x, C_y have already been defined and $C_z = \sum_{z \neq x,y} D(z, A)$. Let's first focus on this latter set, by NEU we must have $xI^{\mathcal{A}}y$. Suppose to the contrary that $xP^{\mathcal{A}}y$ and take a permutation π such that $\pi(x) = y$, $\pi(y) = x$ and $\pi(z) = z$ for all $z \neq x, y$. Then we have $yP^{\mathcal{A}}x$, however the dataset has not changed and therefore we violate the definition of welfare method as a function.

The rest of the proof is by induction on $C_x + C_y$. The inductive base is proved for $C_x + C_y = 2$. Let $C_x + C_y = 1$ and x is chosen; by IR and NEU we have $xP^{\mathcal{A}}y$. If $C_x + C_y = 2$ and $C_x > C_y$ then $xP^{\mathcal{A}}y$ by CNN; if $C_x = C_y$ then $xI^{\mathcal{A}}y$ by NEU. Suppose the statement holds for $C_x + C_y = n$ and we add an observation (x, A). If $C_x - C_y = 1$ then $xP^{\mathcal{A}}y$ by IR and the inductive hypothesis; if $C_x - C_y > 1$ then $xP^{\mathcal{A}}y$ by CNN and the inductive hypothesis; if $C_x = C_y$ then $xI^{\mathcal{A}}y$ by N.

A.2 Proposition 2

By Transitivity, Completeness of R and the finiteness of X; we can make use of a result from Krantz et al. (1971): there exists a real-valued function ϕ on X such that for all $x, y \in X$; xRy if and only if $\phi(x) \geqslant \phi(y)$.

A corollary of this result goes as follows: let $\phi : X \to R^{n-1}$, where $|X| = n$, be a vector valued function and $\phi(x)_z$ be the valued assigned to x when compared to z. Then by the previous result $\phi(x)_z = \phi(x)_y$ for all $y, z \neq x$. The proof is trivial. Suppose the following is false; then we may have $\phi(x)_y > \phi(y)_z > \phi(z)_x$ violating transitivity.

We are now ready to prove our Proposition. Given two generic elements x, y we can partition the dataset in eight disjoint sets with the following cardinalities: C_{xy}, C_{yx} have already been defined; $C_{x,-y} = \sum_{y \notin A} D(x, A)$ and similarly $C_{y,-x}$; $B = B_{xy} = B_{yx} = \sum_{z \neq x,y} \sum_{x,y \in A} D(z, A)$; $D_{xy} = \sum_{z \neq x,y} \sum_{x \in A \& y \notin A} D(z, A)$ and similarly D_{yx}; $E = E_{xy} = E_{yx} = \sum_{z \neq x,y} \sum_{x,y \notin A} D(z, A)$.

Let's first focus on B and E. On these parts of the dataset, NEU implies $xI^D y$. Suppose then $u(x) \geqslant u(y)$, then $C_{x,-y} \geqslant C_{y,-x}$. Using Proposition 1, we have $xR^D y$. Similarly, $C_{xy} \geqslant C_{yx}$ implies $xR^D y$ by IR, CNN and NEU. Note that the premise of our result, namely $u(x) \geqslant u(y)$ implies $C_{x,-y} > C_{y,-x}$ and $C_{xy} > C_{yx}$, doesn't hold for a generic domain D. However, it holds on $\mathtt{hom}(D)$.

To complete the proof we need to extend the argument to D_{xy} and D_{yx}. However, note that $u(x) > u(y)$ implies $D_{yx} > D_{xy}$ and there are no constraints on how such observations should influence the ranking between x, y since a third element is chosen. Hence, a method that attach a positive value on the observations of the type D_{xy}, D_{yx} could led to $u(x) > u(y)$ and $yP^C x$. However, by the corollary of Krantz et al. (1971) result, which is based on Transitivity, we can focus on $D_x = \sum_{z \neq x} \sum_{x \in A} D(z, A)$ instead of D_{xy}. In other words, the value assigned by a method to the observation (z, A) with $x \in A$ and $y \notin A$ must be equal to the one of the observation (y, A) with $x \in A$ and $z \notin A$; otherwise this could potentially lead to cycles. Hence, suppose by contradiction that $u(x) > u(y)$ and $yR^{C_D}x$; then it must be that the value attached to observations in D_x is positive, since $D_y > D_x$. However, we proved that $xR^{C_D}y$ over the parts of the dataset with cardinalities $C_{x,-y}$, $C_{y,-x}$,C_{xy}, C_{yx}, B, E. Suppose we add an observation (x, A) with $y \in A$. Clearly, D_y increase by a positive value. However, since we assumed $yR^{C_D}x$ then CNN is violated.

A.3 Theorem 1

The proof is by induction over the number of observations in the dataset. Denote $|D| = \sum_{A \subseteq X} \sum_{z \in X} D(z, A)$. The induction base is proved for $|D| = 2$. Let $|D| = 0$; by NEU $xI^D y$. If $|D| = 1$ then if $D(z, A) = 1$ by IND we have $xI^D y$; if $D(x, A) = 1$ by SIR we have $xP^D y$. Let $|D| = 2$. If z is chosen the previous result holds by IND. So, let's x or y be chosen. Let the observation (x, A) be added such that $C_x = 2$. By ST we have $\neg yP^{D+(x,A)}x$. Suppose $xI^{D+(x,A)}y$, by SIR we should have $yP^D x$ contradicting the result at $|D| = 1$; hence $xP^{D+(x,A)}y$. If we add an observation (y, A) such that $C_x = C_y = 1$ then by ST and $xP^D y$ we have $\neg yP^{D+(y,A)}x$. Suppose $xP^{D+(y,A)}$ and let (x, B) be the other observation, by ST we should have $\neg yP^{D-(x,B)}x$ violating the result at $|D| = 1$ since $C_y = 1$ and $C_x = 0$. Hence, $xI^{D+(y,A)}y$.

Suppose the result holds for $|D| = n$ and add an observation from a generic set A. Suppose $C_x = C_y$. If $D(z, A) = 1$ then by IND and the inductive hypothesis $xI^{D+(z,A)}y$. If $D(x, A) = 1$, by inductive hypothesis we have $yP^D x$ and by ST $\neg xP^{D+(x,A)}y$. But then since $C_x = C_y$, there exists a set B such that $D(y, B) > 0$. By inductive hypothesis $xP^{D+(x,A)-(y,B)}y$ and by ST $\neg yP^{D+(x,A)}x$. Hence, $xI^{D+(x,A)}y$.

Suppose $C_x > C_y$. If (z, A) the result holds by IND and the inductive hypothesis. If (x, A) then we may have two scenarios: either $xI^D y$ or $xP^D y$. If $xI^D y$ by SIR we have $xP^{D+(x,A)}y$. If $xP^D y$ then by ST $\neg yP^{D+(x,A)}x$. Suppose by contradiction $xI^{D+(x,A)}y$ then by SIR $yP^D x$ contradicting the inductive hypothesis. Hence $xP^{D+(x,A)}y$. If (y, A) then by ST and the inductive hypothesis we have $\neg yP^{D+(y,A)}x$. Since $C_x > C_y$ there exists a set B such that $D(x, B) > 0$; hence suppose by contradiction that $xI^{D+(y,A)}y$; by SIR we have $yP^{D+(y,A)-(x,B)}x$ violating the inductive hypothesis. Hence, $xP^{D+(y,A)}$ completing the proof.

A.4 Theorem 2

The proof of this theorem is very similar to the one of the Theorem 2. We follow the same structure. If $|D| = 0$ we have $xI^D y$ by NEU. Let $|D| = 1$, with one observation from set A. If either $x \notin A$ or $y \notin A$ then by CON $xI^D y$. If $x, y \in A$ and $D(z, A) = 1$ is observed then by NEU $xI^D y$. If $D(x, A) = 1$ and $y \in A$ then by IR we have $xP^D y$. Let $|D| = 2$. If z is chosen, $C_{xy} = C_{yx} = 0$ and x, y are in both sets then by CON and NEU $xI^D y$; if either x or y are not in the set then $xI^D y$ by CON. If $C_{xy} = 1$ then $xP^D y$ by CON and NEU. Let $D(x, A) = 1$ and $C_{xy} = 2$ then by ST $\neg yP^{D+(x,A)}x$; suppose by contradiction $xI^{D+(x,A)}y$, by IR we have $yP^D x$ violating the result at $|D| = 1$; hence $xP^{D+(x,A)}y$. If $C_{xy} = C_{yx} = 1$ then by the argument in Thm 2 using ST we have $xI^{D+(x,A)}y$.

Let $|D| = n$. If (z, A) is the added observation then the result follows by CON, NEU and the inductive hypothesis. If (x, A) is added; let $C_{xy} = C_{yx}$; by the argument in Thm 2 using ST we have $xI^{D+(x,A)}y$. If $C_{xy} - C_{yx} = 1$ then by IR and the inductive hypothesis $xP^{D+(x,A)}y$. If $c_{xy} - C_{yx} > 1$ then by ST and inductive hypothesis $\neg yP^{D+(x,A)}x$. Suppose by contradiction $xI^{D+(x,A)}y$, then by IR $yP^D x$ contradicting the inductive hypothesis. Hence $xP^{D+(x,A)}y$. The argument can be repeated using (y, A) completing the proof.

A.5 Proposition 3

Notice the following trivial fact:

$$d_s(D, P) = \sum_{(x,A)\in O} |\{y \in A : yPx \ \ \& \ \ (x, A)\}| = \sum_{x,y\in X} |\{(x, A) : y \in A \ \ \& \ \ yPx\}|$$

Hence, the number of swaps can be rewritten as:

$$\sum_{x,y\in X} C_{yx} \quad \text{when} \quad xPy$$

In general the maximum number of swaps is: $\sum_{x,y\in X} C_{xy} + C_{yx}$. Define a new measure $\Delta(C, P)$ that equivalently to the swaps distance defines the degree of similarity between a dataset and an irreflexive order P:

$$\Delta(C, P) = \sum_{x,y\in X} [C_{xy} - C_{yx}] \quad \text{when} \quad xPy$$

We prove that for all P_1, P_2 the following holds

$$d_s(C, P_1^*) \leqslant d_s(C, P_2^*) \Leftrightarrow \Delta(C, P_1^*) \geqslant \Delta(C, P_2^*)$$

The proof is algebraic. Note that, given xPy:

$$\sum_{x,y \in X} [C_{xy} + C_{yx}] = \sum_{x,y \in X} [C_{xy} + C_{yx}]$$

$$\underbrace{\sum_{x,y \in X} [C_{xy} - C_{yx}]}_{\Delta(C,P)} - \sum_{x,y \in X} C_{xy} = - \underbrace{\sum_{x,y \in X} C_{yx}}_{d_s(C,P)}$$

Hence, if $d_s(C, P)$ increase by $n \in \mathcal{N}$, then it must be that $\Delta(C, P^*)$ decreases by $2n$.

Denote $\hat{P}_{\mathbf{CRP}}$ the transitive closure of $P_{\mathbf{CRP}}$. We can prove the theorem showing that $P_{\mathbf{CRP}}$ maximizes $\Delta(C, P)$. If $P_{\mathbf{CRP}}$ is acyclic and $xP_{\mathbf{CRP}}zP_{\mathbf{CRP}}y$ and $xI_{\mathbf{CRP}}y$, we have that if $x\hat{P}_{\mathbf{CRP}}y$ then $C_{xy} \geqslant C_{yx}$ for all $x, y \in X$. Hence, $\hat{P}_{\mathbf{CRP}}$ maximize $\Delta(C, P_{\mathbf{CRP}})$. In fact, suppose $yP_{\mathbf{MS}}x$, then by transitivity of $P_{\mathbf{MS}}$, either $zP_{\mathbf{MS}}x$ or $yP_{\mathbf{MS}}z$. Hence, since $C_{xy} = C_{yx}$, $C_{xz} > C_{zx}$ and $C_{zy} > C_{yz}$, we must have that $\Delta(C, P_{\mathbf{MS}}) < \Delta(C, \hat{P}_{\mathbf{CRP}})$, contradicting the definition of $P_{\mathbf{MS}}$.

A.6 Claim 1

We prove that **TC** satisfies IR. The argument follows from Axiom 1, called Prudence, of Nishimura (2017) and the definition of **TC**. The axiom is stated as: $xR^D_{\mathbf{TC}}y$ implies $xR^D_{\mathbf{CRP}}y$. Hence, $xI^D_{\mathbf{TC}}y$ implies $xI^D_{\mathbf{CRP}}y$. The converse is true only if the definition of **TC** holds for all $z \neq x, y$. Then, suppose we add an observation (x, A) with $y \in A$ such that $xP^D_{\mathbf{CRP}}y$. By definition of **TC**, setting $z = y$ we have $yR^D_{\mathbf{CRP}}y$ by reflexivity but $\neg yR^D_{\mathbf{CRP}}x$; hence $\neg yR^D_{\mathbf{TC}}x$. Clearly $xR^D_{\mathbf{CRP}}y$ still holds. Hence, $xP^D_{\mathbf{TC}}y$ and IR is satisfied.

B Independence of the axioms

B.1 Proposition 2

NEU: Suppose **CC** applies to any element apart from y which is always at the bottom of the ranking. This method satifies IR, CNN, T but not NEU.

CNN: Define $xR^{\hom(D)}y$ if and only if $H_{xy} \geqslant H_{yx}$ where $H_{xy} = a \cdot C_{xy} - b \cdot C_{x,-y}$ with $a = \approx 0$. This method satisfied T over $\hom(D)$ when the dataset is the outcome of an i.i.d. RUM. It also

satisfies IR when $a > 0$ and NEU but not CNN. Note that, this method reverse the order defined by the underlying utility of the RUM.

IR: If $xI^{\hom(D)}y$ for all $x, y \in X$; then NEU, CNN and T are satisfied but not IR.

T: Define $F_{xy} \geqslant F_{yx}$ if and only if $xR^{\hom(D)}y$ where $F_{xy} = \delta \cdot C_{xy} + D_{xy}$ with $\delta \in \mathfrak{R}^{++}$. This method satisfies IR, NEU, CNN but not T. In fact, take $\delta \approx 0$ and a Luce Model with $u(x) = 3$, $u(y) = 2$, $u(x) = 1$ and the dataset being 60 observations per each non-empty subsets of X with at least two elements. The reader may note that $F_{xy} < F_{yx}$.

B.2 Theorem 1

ST: Define $N_{xy} = C_{xy} + \delta \sum_{A \not\ni y} D(x, A)$ and $N_{xy} \geqslant N_{yx} \Leftrightarrow xR^D y$. This welfare method satisfies NEU, IND, SIR but not ST.

SIR: The **CRP** method satisfies NEU, IND, ST but not SIR.

IND: Define $T_{xy} = \sum_{A \not\ni y} D(z, A)$ with $z \neq x$ and $C_x + T_{xy} \geqslant C_y + T_{yx} \Leftrightarrow xR^D y$. This welfare method satisfies ST, SIR, NEU but not IND.

NEU: Take a welfare method that ranks $xP^D y$ for all $x \in X$ and all datasets; for all others x, z the **CC** method applies. Note that SIR is satisfied since the antecedent is always false for y. ST, IND are also satisfied while NEU is not.

B.3 Theorem 2

ST: Define $Q_{xy} = \sum_{A \ni x,y} D(x, A) \cdot |A|$ and $Q_{xy} \geqslant Q_{yx} \Leftrightarrow xR^D y$. This welfare method satisfies CON, IR, NEU but not ST.

IR: Let $xI^D y$ for all $x, y \in X$ and all datasets; this welfare method satisfies CON, NEU, ST but not IR; note that ST is satisfied vacuously.

CON: The **CC** method satisfies NEU, ST, IR but not CON.

NEU: Take a welfare method that ranks $xP^D y$ for all $x \in X$ and all datasets; for all others x, z the **CRP** method applies. This welfare method satisfies ST, IR, CON but not NEU.