WZB
Wissenschaftszentrum Berlin
für Sozialforschung

Daniele Caliari

Behavioural Welfare Analysis and Revealed Preference: Theory and Experimental Evidence

Discussion Paper

SP II 2023–303

July 2023

Research Area
Markets and Choice

Research Unit
Economics of Change

Wissenschaftszentrum Berlin für Sozialforschung gGmbH
Reichpietschufer 50
10785 Berlin
Germany
www.wzb.eu

Affiliation of the authors:

Daniele Caliari, WZB (daniele.caliari@wzb.eu)

Abstract

Behavioural Welfare Analysis and Revealed Preference: Theory and Experimental Evidence *

Behavioural welfare economics provides tools to elicit welfare preferences when individuals use nonstandard behavioural models. Current proposals either require assumptions on the models or elicit preferences that become coarser and coarser as the dataset grows. We propose an informational property [Informational Responsiveness] that solves the coarseness problem and, as the dataset grows, characterizes the family of welfare preference elicitation tools that elicit the underlying utility function of a broad family of stochastic models, denoted as preference monotonic models. As such, we argue that Informational Responsiveness is an important property of preference elicitation tools. We then test our property in an experiment in which participants first face a sequence of questions regarding time and risk outcomes and second report their preferences over a subset of the alternatives. We find that preference elicitation tools that satisfy our requirement provide a significantly better match between the elicited and the reported welfare relation.

Keywords: Behavioural Welfare economics, Bounded rationality, Stochastic choice, Revealed preference.

JEL classification: D0, D6

---

* I am indebted to Marco Mariotti and Christopher Tyson for the numerous pieces of advice. I also thank Jose Apesteguia, Miguel Ballester, Dhruva Bhaskar, Miguel Costa-Gomes, Dirk Engelmann, Elisa Facchetti, David Freeman, Georgios Gerasimou, Sean Horan, Steffen Huck, Maria Vittoria Levati, Aniol Llorente-Saguer, Lorenzo Neri, Thomas Palfrey, Ludovic Renou, Vasiliki Skreta, Ivan Soraperra, and seminar participants at Queen Mary University internal seminars, CEPET Workshop, RES 2019 Annual Conference, EEA-ESEM 2019 Summer Congress, 2019 Summer School of the Econometric Society, 6th LSE-Oxford Graduate Student Conference, ES 2020 World Congress, ESA 2020 Job Market Seminar Series. I gratefully acknowledge financial support from the Royal Economic Society and Queen Mary College.
Market Behaviour and Economics of Change Research Unit, Social Science Research Center Berlin (WZB). E-mail: daniele.caliari@wzb.eu

All discussion papers are downloadable:
http://www.wzb.eu/en/publications/discussion-papers/markets-and-choice

1 Introduction

In recent years, behavioural economics has developed a large number of behavioural models in response to evidence of violations of the standard model of decision-making. This growing literature raises the problem of selecting behavioural models to analyse datasets and contribute to policy evaluation. More concretely, imagine a researcher who collects data to infer welfare preference relations. She does not observe individuals' choice procedures but only their final choices. In this scenario, some individuals may behave as utility maximizers. However, others may not. Some may face costs of thinking (Ortoleva (2013), Fudenberg et al. (2015), Frick (2016)), form consideration sets (Manzini & Mariotti (2014a), Brady & Rehbeck (2016), Caplin et al. (2019)), use attention filters (Masatlioglu et al. (2012), Lleras et al. (2017), Cattaneo et al. (2020)), perception orders (Echenique et al., 2018), checklists (Mandler et al., 2012) or sequential rationales (Manzini & Mariotti, 2007).

The literature has acknowledged the complexity of the researcher's task. Chetty et al. (2009) define it as: "... the key challenge for behavioural welfare economics." To tackle this problem, Bernheim & Rangel (2009) propose an approach that can be applied irrespective of individuals' behavioural models. The authors develop a Pareto relation that cautiously regards **x** as preferred to **y** if and only if **y** is never chosen when **x** is available. The researcher's task is simplified, as she no longer needs to identify the different models anymore. However, this simplification comes at a considerable cost: the elicited preference may be very coarse. This problem has been highlighted by Rubinstein & Salant (2012) as follows: "The resulting Pareto relation is typically a coarse binary relation that becomes even more so as the behavioural dataset grows." For instance, imagine a utility maximizer that at the act of choice commits a mistake with a small probability. As Rubinstein & Salant (2012) noticed, more observed choices imply a higher chance of eliciting an empty Pareto relation.

The described drawback of Bernheim & Rangel's (2009) approach is informational in its nature as noticed by Manzini & Mariotti (2014b): "... some choice situations which are 'suspect' may nevertheless provide information about the decision mechanism used by the agent when 'crossed' with non-dubious choices". The authors raise the case for model-based approaches to the researcher's problem, in contrast to model-free (or model-less) approaches (Bernheim & Rangel, 2009). We claim that there is no such thing as model-free approaches, while the distinction is between explicitly and implicitly model-based approaches. The former relies on specific models of decision-making with specific revealed preference mappings. The latter is based on assumptions on the revealed preference mappings that only implicitly pose constraints on the models of decision-making. Henceforth, we generically refer to mappings from individual choices to welfare preference relations as welfare methods (or simply methods).

In this paper, we claim that a solution to the informational (or coarseness) problem raised by Rubinstein & Salant (2012) and Manzini & Mariotti (2014b) does not necessarily require an explicitly model-based approach. We propose a normative principle for welfare methods, called Informational Responsiveness, that guarantees that "more observations lead to finer results". Informational Responsiveness states that when the researcher regards **x** as indifferent to **y**, then more choices of **x** with **y** available should turn the judgement in favour of **x**.[1] A violation would imply that she discards these observations. Beyond the informational interpretation, Informational Responsiveness implies that observations in which **x** is chosen when **y** is available, or vice versa, have to be welfare-relevant, hence it selects a particular notion of frequency. We investigate further this side of Informational Responsiveness - the frequency interpretation[2] - studying the implications of Informational Responsiveness when observations in which **x** is chosen when **y** is available, or vice versa, are the *only* welfare-relevant observations, as in standard revealed preference theory (Arrow (1959), Sen (1971)).

We argue that Informational Responsiveness is an important condition for a reliable solution to the researcher's complex problem. We propose both theoretical and experimental evidence to corroborate this claim. In Propositions 1 and 2, we argue in favour of the informational interpretation of Informational Responsiveness as we show that, as the behavioural dataset grows, welfare methods that fail to satisfy it elicit coarser and coarser welfare relations as they waste more and more data points. In Proposition 3, we combine the informational and frequency interpretations to characterize a large family of welfare methods. We show that our property implicitly selects the broad family of

[1] To the best of our knowledge, this property was first introduced in voting theory by Goodin & List (2006) under the denomination of "One Vote Responsiveness".

[2] Bernheim & Rangel (2009) discuss the coarseness problem and some possible solutions in section VII of their paper. We particularly focus on their section VII.C. It is from this section that we borrow the term "frequency" as the authors, who mentioned our approach as "preponderance criterion", highlight as a conceptual problem the fact that "there are potentially many competing notions of frequency". Here, we study the most classical notion of frequency that in line with standard revealed preference theory.

decision-making models in which an element **x** is preferred to **y** if and only if whenever **x** and **y** are available, **x** is more likely to be chosen than **y**. We denote these models as *preference-monotonic* models. This result is a generalization of Apesteguia & Ballester (2015)[Proposition 1 & Theorem 1] in which the authors show that their welfare method, shortly denoted as Minimum Swaps,[3] is the solution to the researcher's problem when individuals adopt specific models such as i.i.d. Random Utility models, Tremble models or Additive Perturbed Utility models (Fudenberg et al., 2015).[4] Our generalization is in two directions: first, we show that there is a broad family of welfare methods that can solve the researcher's problem under the above set of models; second, we show that the result is achievable under much weaker conditions, encompassing models that may violate standard stochastic properties such as regularity and strong (or moderate) stochastic transitivity (Fudenberg et al., 2014). Finally, our family of welfare methods has a natural representation under preference-monotonic models. Namely, **x** and **y** are evaluated based on the frequency in which **x** is chosen when **y** is available as vice versa. Such frequencies can receive different weights depending on the sets in which they are observed with the only constraint that the weights must be strictly positive. This simple representation will allow us in the experimental part of the paper to construct a clean and direct test of Informational Responsiveness.

In the second part of the paper, we use a novel choice elicitation experiment to recreate the researcher's problem. Subjects are asked to choose from sets that include delayed payment plans (Time Preference) or lotteries (Risk Preference). The task of the researcher is to elicit welfare relations based solely on choices.

First, as the premise of our theoretical and empirical analysis, we find that 63% of participants violate the Weak Axiom of Revealed Preference in Time and 94% violate it in Risk. This latter finding is in line with Agranov & Ortoleva (2017). Therefore, we confirm the necessity of welfare methods that allow us to elicit reliable welfare relations for a substantial fraction of the subjects for whom standard methods do not apply.

To evaluate welfare methods, following an approach in line with (ordinal) liking-

[3]This welfare method assigns to a dataset of choices the preference that minimizes the sum over all observations of the number of alternatives that are ranked above the chosen one.

[4]More specifically, Apesteguia & Ballester (2015) define the property of P-Monotonicity which states that if a decision maker prefers **x** to **y**, then **x** is chosen with higher probability than **y** when both are available. They show that the preference relation implied by P-Monotonicity is the one elicited by Minimum Swaps [Theorem 1] and that the above stochastic models all satisfy P-Monotonicity [Proposition 1].

rating tasks (Reutskaja et al., 2011), at the end of the experiment, we ask participants to rank four alternatives in both the risk and time environments. In an exercise that we call "Identification", we measure the proportion of subjects for whom each welfare method can elicit either the entire reported preference relation or simply the reported best alternative. The problem of evaluating welfare methods does not have a clear answer as it relies on the researcher's assumptions on the behavioural models or on the welfare-relevant decisions as expressed in Bernheim & Taubinsky (2018). Previous papers (e.g. Manzini et al. (2010), Bouacida & Martin (2021)), focused on the properties of the elicited welfare relations. However, first, different welfare methods have different ex-ante theoretical properties, making an analysis of ex-post properties a biased indicator of their efficacy. Second, the alternatives analysed were monotonically ranked. This led to a very low number of inconsistencies which did not challenge the limitations of Bernheim & Rangel's (2009) approach (Bouacida & Martin, 2021).

In line with our theoretical predictions, as an indirect test of Informational Responsiveness, we find that welfare methods that satisfy it outperform the other methods. When asked to uniquely identify the best reported alternative, the Pareto approach (Bernheim & Rangel, 2009) is outperformed by 32% in Time and 47% in Risk. Similarly, when asked to uniquely identify the entire welfare preference relation, the Pareto approach is outperformed by 20% in Time and 18% in Risk.[5] These results are robust, and hold when: (1) we allow for the set identification of the best alternative; (2) we discriminate methods using a measure of similarity between the reported and elicited welfare relation; (3) we enlarge our comparison to a wide variety of welfare methods that do and do not satisfy Informational Responsiveness.

A possible caveat of our analysis is related to the possible misalignments between

[5]These percentages are calculated on the total number of subjects. The Pareto relation uniquely identifies the correct best alternative of 55% of the subjects in Time and 42% in Risk, while the Minimum Swaps method of 87% in Time and 62% in Risk. Focusing on the entire welfare relation, the Pareto relation uniquely identifies it for 42% of the subjects in Time and 6% in Risk, while the Minimum Swaps method for 62% of the subjects in Time and 24% in Risk.

choices and reported preferences.[6] We address this problem in two ways. First, our set identification results provide a measure of alignment between the elicited and the reported welfare relation. Second, following Fudenberg et al. (2022), we use machine-learning techniques to estimate a measure of irreducible error - namely, the proportion of subjects who cannot be identified by any method, as they fail to be identified by a data-driven optimal weighting algorithm. In doing so, we provide a general measure of performance for the welfare methods. We confirm that the methods that satisfy Informational Responsiveness are not only significantly more complete but also strikingly close to the data-driven algorithm.

As a final contribution and as previously mentioned, we propose a direct test of Informational Responsiveness. We use the data-driven algorithm to test whether all types of sets convey information regarding the welfare of the subjects. Our representation result in Proposition 3 implies that Informational Responsiveness is satisfied only if the algorithm assigns strictly positive weights to each parts of the dataset. We find that, in Time Preference, the only sets that fail to satisfy Informational Responsiveness are those characterized by asymmetric dominance relations, and therefore potential attraction effects (Huber et al. (1982), Natenzon (2019)). The remaining types of sets are assigned strictly positive weights regardless of their cardinality. The importance of Informational Responsiveness is even more striking in Risk Preference where every type of decision problem is assigned a strictly positive weight, hence regarded as informationally relevant to elicit the (entire) welfare relation of the subjects.

1.1 Related Literature

The theoretical part of the paper is related to the literature on choice-theoretic welfare analysis: Green & Hojman (2007), Salant & Rubinstein (2008), Bernheim & Rangel (2009), Masatlioglu et al. (2012), Rubinstein & Salant (2012), Manzini & Mariotti (2014b), Apesteguia & Ballester (2015), Horan & Sprumont (2016), Nishimura (2018).

[6]Our liking-rating tasks are not incentivized. Hackethal et al. (2022) find no differences between behaviour in incentivized and non-incentivized tasks in a preference elicitation experiment on risk preferences. A similar result was found in Holt & Laury (2002). Enke et al. (2021) show the effect of typical experimental stakes over tasks that involve either correct or incorrect answers, such as the Frederick test (Frederick, 2005). They find no significant effect. The absence of an effect of stakes is reported also in Enke & Zimmermann (2019) in a belief formation setting. Incentivized liking-rating tasks do not solve the issue regarding the possible misalignment between choices and reported preferences, but their introduction may provide interesting insights. Importantly, our paper provides tools to measure and evaluate this misalignment independently from the incentives.

The experimental part first relates to the few existing choice elicitation experiments, such as Manzini et al. (2010), Barberá & Neme (2016) and McCausland et al. (2020). Our experiment differs from these in two main ways: (i) we collect choices on a much richer set of questions to test how behavioural effects, such as asymmetric dominance (Huber et al. (1982), Natenzon (2019)) and choice overload (Iyengar & Kamenica, 2010), influence welfare revelation; (ii) we ask subjects to directly report their preference relation. Secondly, our experimental design relates to the empirical literature on stochastic choice and choice deferral. Although our design shares some features with existent experiments, none of those elicits both choices and preferences, is based on both time and risk preferences, and collects choices from all non-empty subsets of the MAIN alternatives as well as sets with behavioural effects.[7]

1.2 Structure of the paper

The paper's structure is as follows: section 2 introduces the framework and presents the theoretical results. We also describe the welfare methods that will be analysed subsequently. Section 3 presents in detail the experimental design and the hypotheses. The main experimental results are presented in section 4; all of them are divided with respect to Time and Risk. The Appendix contains details regarding the theoretical proofs and further theoretical results. More details regarding the experimental design and further experimental results 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 $\mathcal{O}$ as the set of all possible pairs (x, A) where $A \subseteq X$ and $x \in A$. A dataset $D : \mathcal{O} \to \mathbb{Z}_+$ assigns a non-negative integer to each pair. For instance, we write $D(x, A) = 1$ to say that x has been chosen from A one time. We denote $\mathcal{D}$ as the set of all possible datasets. Note that, this definition of datasets is very general as it allows for multiple observations [$D(x, A) > 1$] and missing data [$D(x, A) = 0$ for all $x \in A$].

[7] Some previous studies 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 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 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.

Denote as $\mathcal{R}(X)$ the set of all complete[8] binary relations, with I and P denoting the symmetric and asymmetric part respectively. 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 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, as well as the Pareto relation (Bernheim & Rangel, 2009). The simple counting, denoted C_x, and the counting revealed preference relations, denoted C_{xy}.

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

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

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

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

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

Having defined the **CRP** method, the definition of the Pareto Relation, or henceforth **BR** method, is straightforward. Bernheim & Rangel (2009) proposed the following idea: 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 (Bernheim & Rangel, 2009, Theorem 1). Formally, $xP^D_{\mathbf{BR}} y$ if and only if $C_{xy} > 0$ and $C_{yx} = 0$. Otherwise, $xI^D_{\mathbf{BR}} y$.

2.1 Informational Responsiveness

We start by providing the context in which our main idea arises and we do so by splitting the well-known property of Positive Responsiveness (May (1952), Rubinstein (1980)),

[8] We do not impose neither acyclicity nor transitivity. Acyclicity is particularly appealing because it guarantees the existence of a maximal element for all $A \subseteq X$ (Sen, 1997). We relax this assumption to define the counting revealed preference procedure, denoted as **CRP**, as a welfare method. Its inclusion is driven by two rationales: (1) **CRP** constitutes the foundation for other welfare methods and (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. binary relations on X.

which in our context can be written 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]. This seemingly standard axiom is strong because it captures simultaneously four different concepts: (1) when the antecedent is concerned with I, it has an informational interpretation given by the ability of the welfare method to break the indifference relation using choices; (2) it has an obvious monotonic interpretation, namely, choices are strictly positive signals of welfare; (3) it has a particular frequency interpretation, namely any choice of x, even when y is not available, is relevant for determining the welfare relation between x and y; (4) it considers one observation as sufficient to break the indifference relation; hence, forcing "thin" indifference classes on the welfare relation.

We represent these concepts using four different axioms to which we add a standard neutrality condition. The first, denoted Informational Responsiveness, henceforth IR, is the main axiom of the paper. Note that all axioms hold for all $A \subseteq X$, $D \in \mathcal{D}$ and for all $x, y \in X$, with $x \neq y$ whenever strict preferences are involved.

Axiom 1 (Informational Responsiveness **[IR]**[9]**).**

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

IR has a double nature, represented by its two antecedents. First, it deals specifically with the informational interpretation of breaking the indifference relation using choices. Second, focusing only on sets in which both x, y are available, it deems as welfare-relevant a particular definition of frequency that is reminiscent of the standard revealed preference analysis (Arrow (1959), Sen (1971)). To formalize the use of the terminology "informational" and "frequency" interpretations of IR we introduce the following two axioms.

Axiom 2 (Connection **[CON]**)**.**

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

Axiom 3 (Strong Informational Responsiveness **[SIR]**)**.**

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

[9]The consequent of this axiom is technically incomplete. We should define it when we both add and remove observations. The complete version is $xP^{D+(x,A)}y$ and $yP^{D-(x,A)}x$ whenever (x, A) is in the dataset D. However, this addition becomes irrelevant when Axiom 4 (**CNN**) applies.

CON is an independence axiom, similar to Arrow's Independence from Irrelevant Alternatives. It states that sets in which two alternatives are not available together should not play any role in shaping the welfare relation between them. SIR, instead, is the relative complement of CON in IR.[10] Namely, it enforces the informational interpretation of IR on sets in which x, y are not necessarily available together.

We define the informational and frequency interpretations of IR as follows: the former is captured by IR singularly, while the latter is captured by IR when the independence axiom CON is satisfied.

Importantly, the introduced welfare methods are perfectly separated by these three axioms. The **CC** method satisfies SIR but not CON, the **BR** method satisfies CON but not IR, while the **CRP** method satisfies IR and CON but not SIR. Therefore, an indirect test of both the informational and frequency interpretations of IR can be constructed by comparing the performances of these three welfare methods.

Among the concepts captured by Positive Responsiveness, IR retains the, perhaps strong, assumption that a single observation is decisive to break the indifference relation. In Appendix B, we consider the possibility of a more cautious approach in which more than one observation is required to trigger a strict welfare judgment. Our theoretical conclusions are not substantially affected. In particular, the informational interpretation of IR and its role of critique to Bernheim and Rangel's (2009) approach are unchanged.

The next axiom covers the most uncontroversial property of Positive Responsiveness, weak monotonicity. Choice Non-Negativeness, henceforth CNN, states that choices are non-negative signals of welfare and, contrarily to IR, it is satisfied by all (implicitly model-based) methods in the literature.

Axiom 4 (Choice non-negativeness **[CNN]).**

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

Finally, Neutrality, henceforth NEU, asserts that a welfare method cannot, a priori, favour or punish some alternatives over others. Since our theory does not rely on any additional information about alternatives or models, NEU is a reasonable assumption.

[10]In fact, on can show that SIR and CON are the exact axioms that distinguish the **CC** and **CRP** methods when characterized over a general dataset and without assumptions on the decision models that have generated the data. This problem, though interesting, is out of the scope of the paper, but details are available upon request.

In fact, similar to CNN, this axiom is satisfied by all methods proposed by the literature and compatible with our framework. Let $\Pi(X)$ be the set of all the permutations $\pi : X \to X$. For all π:

Axiom 5 (Neutrality **[NEU]**).

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

2.2 The role of Informational Responsiveness

In this section, we provide theoretical support to the claim that IR is an important and desirable property of welfare methods. We do so in three steps. First, we show that the informational interpretation of IR is crucial to avoid two paradoxical situations: (i) indisputable preferences are not identified, and (ii) the welfare relation becomes coarser and coarser when the number of observations increases. Second, we show that the informational interpretation alone guarantees this result only under restrictive conditions in terms of behavioural models and structure of the dataset. Finally, we show that imposing the frequency interpretation of IR through the conjunction with CON allows us to restore the result under a wide family of behavioural models and for general datasets.

A common component of behavioural models, both deterministic and stochastic, is a transitive and complete welfare (or preference) relation $\succeq$. We restrict our attention to a broad family of stochastic models characterized by the following monotonic property: for all $x, y \in X$, and for all sets $A \subseteq X$ such that $x, y \in A$, $x \succeq y$ if and only if $p(x, A) \geq p(y, A)$, where $p(x, A)$ is interpreted as the probability of choosing x from the set A. We call this set of models *preference-monotonic* models. We denote them as $\mathcal{MON}$, and we denote the probability of choosing x from a set A under these models as $p_{\mathbf{MON}}(x, A)$.[11]

To translate the use of stochastic models into our discrete definition of dataset we follow Apesteguia & Ballester (2015) and introduce an alignment property. The frequency at which x is chosen from A is given by $\texttt{fr}_{\mathbf{MON}}(x, A) = \rho(A)p_{\mathbf{MON}}(x, A)$, where $\rho(A)$ is the probability of facing the menu A in the dataset. In our discrete setting, we denote N the number of observations in the entire dataset, and $n(A)$ the number

[11]A representation of the set of $\mathcal{MON}$ models has been given by Fudenberg et al. (2014) using a utility function $u : X \to \mathbb{R}_{++}$ and a menu-specific cost function $c_A : [0, 1] \to \mathbb{R}$. Subjects choose the probability distribution over each set A that maximizes the sum of expected utility and the convex cost function c_A. For a technical definition, we refer the reader to Fudenberg et al. (2014). The authors rely on a property called Item Acyclicity which is equivalent to our monotonic property under full-observability of the menus.

of observations on a set A. We say that a dataset is aligned with a model **m** if for all, not necessarily distinct, sets A, B and alternatives $x \in A$, $y \in B$, $\frac{D(x,A)}{n(A)} \geq \frac{D(y,B)}{n(B)}$ if and only if $\frac{\mathtt{fr}_{\mathbf{m}}(x,A)}{\rho(A)} \geq \frac{\mathtt{fr}_{\mathbf{m}}(y,B)}{\rho(B)}$. Importantly, this is true as $N \to \infty$, and as $n(A) \to \infty$ uniformly for all sets $A \subseteq X$. Hence, as the dataset grows it aligns with the underlying stochastic model and allows us to formalize the intuition of Rubinstein & Salant (2012). The alignment property also rules out datasets in which we have no observations about the alternatives under scrutiny.

We start by focusing on the very narrow type of datasets characterized by multiple observations on a single set A; we denote these datasets as $\mathcal{A}$. We show that, given this particular restriction, IR, CNN, and NEU characterize welfare methods that can correctly identify the underlying welfare relation.

Proposition 1. *A welfare method g satisfies IR, NEU, and CNN if and only if* $\succeq = R_g^D$ *for all datasets* $D \in \mathcal{A}$ *aligned with* $\boldsymbol{m} \in \mathcal{MON}$.

Proof. See Appendix A.1. □

Remark 1. Proposition 1 has two immediate corollaries. First, the **BR** method fails to elicit the underlying welfare relation more and more often as N grows. This is the exact translation of the critique of Rubinstein & Salant (2012). Second, the characterized welfare methods elicit the underlying welfare relation more and more often as N grows. An exact translation of the principle "more data lead to finer results".

One important comment; the proof relies on the characterization of the counting procedures **CC** and **CRP**, which are equivalent on datasets $\mathcal{A}$. However, the restriction to $\mathcal{A} \subset \mathcal{D}$ is extremely severe. The equivalence between **CC** and **CRP** can be proved on a larger set of datasets but at a considerable cost. First, a weaker, but still severe, restriction on the datasets has to be maintained. Particularly, a dataset $D \in \mathcal{D}$ is *homogeneous*, denoted as $\mathtt{hom}(D)$, if any $A, B \subseteq X$ with the same cardinality is observed the same number of times, namely $n(A) = n(B)$ if $|A| = |B|$.[12] Second, the resulting welfare relation is required to be transitive. Third, the collection of observations is the result of a smaller set of models. Following Fudenberg et al. (2015), we refer to this set of models as Additive Perturbed Utility models and denote them as $\mathcal{APU}$.

[12]Notice that as the dataset grows the distinction between homogenoeus and non-homogeneous datasets becomes redundant.

In $\mathtt{hom}(D)$, $\mathcal{APU}$ models guarantee not only that the preference relation ranks the alternatives according to their choice frequencies within each set, but that this ranking is preserved between sets.[13] A noticeable member of the class of $\mathcal{APU}$ models is the Luce Model (Luce, 1959), for which the **BR** method would fail to elicit the underlying welfare relation.

Proposition 2. *A welfare method g satisfies IR, NEU, CNN, and Transitivity if and only if* $\succeq = R_g^D$ *for all datasets* $\mathtt{hom}(D)$ *aligned with* $\boldsymbol{m} \in \mathcal{APU}$.

Proof. See Appendix A.2.[14] □

On one hand, Proposition 2 shows that the implications of the informational interpretation of IR alone can be extended to important dataset structures such as the power set or the set of binary sets. On the other hand, and more importantly, it highlights the importance of the frequency interpretation of IR. As a final result, we exploit CON to provide a complete characterization on general datasets of the welfare methods that allow us to elicit the underlying preference relation over the set of models $\mathcal{MON}$. Proposition 3 provides also a representation for the class of methods characterized by IR, NEU, CON, and CNN on datasets resulting from the set of models $\mathcal{MON}$. This representation is based on the **CRP** method, but allows the observations to be weighted differently depending on the menu under scrutiny with the constraint that weights have to be strictly positive. We denote this class as weighted **CRP** methods, or shortly **WCRP**$^+$. We say $xR^D_{\mathbf{WCRP}}y$ if and only if there exists a system of weights $\langle w_A \rangle_{A \subseteq X}$ such that $\sum_{A \subseteq X} w_A C^A_{xy}$ where $C^A_{xy} = D(x, A)$ for all $A \ni x, y$. **WCRP** becomes **WCRP**$^+$ if $w(A) > 0$ for all $A \subseteq X$. This representation is an important link to our experimental analysis. In section 4.3, we exploit the strict positivity of the system of weights to test directly IR. The intuition is simple: if weights are only weakly positive our representation satisfies NEU, CON, CNN, but not IR; while if negative weights are allowed then also CNN is violated.

Proposition 3. *The following are equivalent:*

[13]This observation is a corollary of Theorem 1 in Fudenberg et al. (2015), as these models satisfy a form of Independence from Irrelevant Alternatives, denoted by Fudenberg et al. (2015) as Ordinal IIA.

[14]The importance of transitivity is readily seen. The axioms do not impose constraints on observations (z, A) with $z \neq x, y$, $x \in A$ and $y \notin A$. If these observations are considered by the method, we could observe $u(x) \geq u(y)$ and $yP_g^{\mathtt{hom}(D)}x$. Transitivity excludes these cases.

(i) a welfare method g satisfies IR, NEU, CON, and CNN.

(ii) $\succeq = R_g^D$ *for all datasets* $D \in \mathcal{D}$ *aligned with* $\boldsymbol{m} \in \mathcal{MON}$.

(iii) $g = \boldsymbol{WCRP}^+$.

Proof. See Appendix A.3. □

2.3 Welfare methods

The welfare methods that we analyse empirically are the following: **SEQ**[15] 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(BR)** and **TC(CRP)** are two versions of the *transitive core method* - Nishimura (2018), and **ATT(BR)** and **ATT(MS)** are two versions of the *model-based attention method* - Masatlioglu et al. (2012), Cattaneo et al. (2020). Here, we briefly introduce them and, if needed, provide results regarding their connection with our theoretical results. In Figure 1 (section 3.1), We discuss our experimental hypotheses and summarize the properties of these methods.

2.3.1 Sequential

The sequential method (Horan & Sprumont, 2016) can be effectively applied only if the dataset is constrained on $\mathcal{X}$ and it has one observation for each set.[16] 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, we write $xP^D_{\mathbf{SEQ}}y$ for all $y \neq x$, if $D(x, X) = 1$; $yP^D_{\mathbf{SEQ}}z$ for all $z \neq x, y$ if $D(y, X \setminus \{x\}) = 1$; $zP_{\mathbf{SEQ}}w$ for all $w \neq x, y, z$, if $D(z, X \setminus \{x, y\}) = 1$ and so on.

2.3.2 Minimum swaps

Apesteguia & Ballester (2015) proposed a swaps index,[17] defined as the sum over all

[15]This method has been indirectly proposed by Aleskerov et al. (2007)[Theorem 5.5] to construct a utility function from choices that satisfy a property denoted as Fixed Point.

[16]Horan & Sprumont (2016) suggest a way to extend **SEQ** over different datasets simply by taking the intersection of all resulting orderings. However, we do not apply this extension empirically because it would not provide any additional information to the analysis of this method.

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

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

observations of the number of alternatives that are ranked above the chosen one. The preference relation P that minimizes this index is the solution of the minimum swaps method, denoted as **MS**. When more than one preference relation P minimize the above problem, we adopt the convention of taking the intersection among all the minimizers.

This method is of particular interest to us because of its strict connection with **CRP**. In Proposition 4 below, 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 $P^D_{\mathbf{MS}}$. Furthermore, Propositions 3 and 4 together generalize Apesteguia & Ballester (2015)[Proposition 1 & Theorem 1] and imply that, when the dataset is the result of the set of models $\mathcal{MON}$, **MS** is equivalent to **CRP**. The reader may note that **MS** does not generally satisfy CON, however, violations of CON can only arise if $P^D_{\mathbf{CRP}}$ is cyclic, which is impossible under the assumptions of Proposition 3.[18]

Let P^* be the transitive closure of P, namely the smallest relation P^* that contains P and it is transitive.

Proposition 4. *If $P^D_{\mathbf{CRP}}$ is acyclic, then $xP^{*D}_{\mathbf{CRP}}y \Leftrightarrow xP^D_{\mathbf{MS}}y$.*

Proof. See Appendix A.4. □

2.3.3 Eigenvector centrality

This novel method uses the definition of centrality in networks to define an order of alternatives. First, we construct the weighted revealed preference graph using C_{xy}. The adjacency matrix $A = (\omega_{xy})_{x,y\in X}$ is defined as follows:

$$\omega_{xy} = \begin{cases} C_{xy} & \textbf{if} \quad C_{xy} > 0 \\ \varepsilon & \textbf{if} \quad C_{xy} = 0 \end{cases}$$

given a small $\varepsilon > 0$. The eigenvector centrality of the nodes in the graph induces a complete and transitive welfare relation that measures the importance of each alternative.[19] Differently from **MS**, Proposition 3 does not fully extend to **EIG**. Particularly, under

[18] Proposition 4 is in fact stronger since it proves the equivalence between **CRP** and **MS** more generally than under the conditions imposed in Proposition 3.

[19] The eigenvector centrality of $x \in X$, denoted as c^e_x, is:

$$c^e_x = \frac{1}{\lambda_{max}} \sum_{y\in X} \omega_{xy} c^e_y$$

Its existence is guaranteed by Perron-Frobenius Theorem. Simply, for all $x, y \in X$ we have $xR_{\mathbf{EIG}}y$ if and only if $c^e_x \geq c^e_y$.

the conditions of Proposition 2, **EIG** guarantees preference elicitation, while this may not be the case under the conditions of Proposition 3.[20]

2.3.4 Transitive core

Nishimura (2018) recently proposed a methodology that aims to infer the welfare relation from a complete but not necessarily transitive revealed preference relation. We analyse his original proposal, denoted **TC(BR)**, which was in line with Bernheim & Rangel (2009),[21] and a variation based on **CRP** and denoted **TC(CRP)**. The former does not satisfy IR, while the latter does. The transitive core method for $\mathbf{i} \in \{\mathbf{BR}, \mathbf{CRP}\}$ is defined as follows:

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

It is trivial to notice that Proposition 3 extends to **TC(CRP)** since if $R^D_{\mathbf{CRP}}$ is transitive and complete then $R^D_{\mathbf{CRP}} = R^D_{\mathbf{TC(CRP)}}$. This is also an implication of the Axiom 2 (denoted "Principle of revealed preferences") by Nishimura (2018). However, note that the alignment between $P_{\mathbf{CRP}}$ and $P_{\mathbf{TC(CRP)}}$ does not extend to acyclicity as in **MS**. For instance, if $xP_{\mathbf{CRP}}yP_{\mathbf{CRP}}z$ but $xI_{\mathbf{CRP}}z$, one can easily see that $xP_{\mathbf{TC(CRP)}}z$, and $xI_{\mathbf{CRP}}yI_{\mathbf{CRP}}z$.[22]

2.3.5 Explicitly model-based attention approach

As last, we analyse the explicitly model-based method of Masatlioglu et al. (2012), denoted **ATT**, which is notably in contrast with IR. We briefly introduce the model and the implied welfare relation. A decision-maker is endowed with a complete and transitive preference relation and an attention filter.[23] When facing a set, the decision-maker first forms a consideration set using the attention filter, and then selects the best element according to the preference relation. Masatlioglu et al. (2012)[Theorem 1] shows that the welfare relation of this attention model [$R^D_{\mathbf{ATT}}$] is the transitive closure of the following

[20]The counterexample proposed at the end of Appendix A.2 shows that under models $\mathcal{MON}$ we may have a contradiction between **CRP** and **EIG**.

[21]The author wrote: "We may observe the decision maker's choices from the same choice problem on more than one occasion. If she chooses one alternative on some occasions and another on others, then we reveal indifference between these two alternatives". (Nishimura, 2018)

[22]The transitive core relation has been completed since one can easily check that, for instance, $\neg xR_{\mathbf{CRP}}y$ and $\neg yR_{\mathbf{CRP}}x$.

[23]An attention filter is a mapping $\Gamma : \mathcal{X} \to \mathcal{X}$ such that for any $A \in \mathcal{X}$: $\Gamma(A) \subseteq A$ and $\Gamma(A) = \Gamma(A \setminus x)$ whenever $x \notin \Gamma(A)$.

relation: xPy if there exists a set A such that the decision-maker chooses x from A but a different element from $A \setminus y$. To experimentally test this method, as well as for theoretical considerations, we need to overcome two difficulties. The first, acknowledged by the authors, is that $R^D_{\mathbf{ATT}}$ is a very coarse relation because it relies on violations of Sen's Property α[24] to be completed. As the authors wrote: "[the welfare relation] is empty if the choice data satisfies WARP." To solve this issue, we follow closely the suggestions of the authors. Namely, we first elicit $R^D_{\mathbf{ATT}}$ and then we complete it using either the **BR** or the **MS** methods. The resulting welfare relations may contain cycles due to the disagreements between the methods. In such cases, as suggested by the authors, the cycles are broken favouring $R^D_{\mathbf{ATT}}$. Finally, if the choices are incompatible with the attention model, we simply rely on **BR** and **MS**. The results are two new welfare methods **ATT(BR)** and **ATT(MS)**.

The second difficulty relates to the fact that the welfare relation $R^D_{\mathbf{ATT}}$ is based on a deterministic model, while our theoretical framework deals with possible multiple observations at each set. To overcome this drawback, we build on Cattaneo et al. (2020). Here, the authors introduce a stochastic model called Random Attention Model (RAM) and, in Theorem 1, they characterize its revealed preference relation. Particularly, in RAM, x is revealed preferred to y if and only if $p(x, A) \geq p(x, A \setminus y)$. This welfare relation can be readily translated into our framework as follows: $xR^D_{\mathbf{ATT}}y$ if and only if there exists a set $A \ni y$ such that $D(x, A) \geq D(x, A \setminus y)$. In RAM, the existence of a welfare relation is connected to a violation of Regularity (Marschak & Block, 1960), a property of stochastic models that states that the probability of choosing an alternative is decreasing in set inclusion. We leave it to the reader to check that if a stochastic model yields a degenerate probability, Regularity is equivalent to Sen's Property α. In this sense, we interpret $R^D_{\mathbf{ATT}}$ as the welfare relation of both the deterministic model of limited attention (Masatlioglu et al., 2012), and the RAM (Cattaneo et al., 2020).

Few final remarks. Both **ATT**, **ATT(BR)**, and **ATT(MS)** violate IR. However, in the first two cases, the violation arises because both **ATT** and **BR** violate IR, while in the last case, it arises because **ATT** is considered lexicographically more important than **MS** in case of disagreement. Finally, given the extreme coarseness of **ATT**, in the experimental part of the paper, we focus solely on **ATT(BR)** and **ATT(MS)**, and

[24]Sen's property α states that if an element is chosen from a set A then it must be chosen from all its subsets.

given the particular structure of $R^D_{\mathbf{ATT}}$ we apply these methods only if the dataset is constrained on $\mathcal{X}$, as for the sequential method.

3 Experimental design

The experiment follows a standard choice elicitation design, e.g. Manzini et al. (2010), Barberá & Neme (2016). The complete instructions and screenshots are presented in the Online Appendix. Participants received instructions both on screen and on paper such that they could consult them during the experiment.

The experiment is divided into three parts: (1) choice elicitation, (2) questionnaire, and (3) Raven Test. The choice elicitation part has 50 questions: half have to do with choices among lotteries (Risk Preference Elicitation) and half with choices among delayed payment plans (Time Preference Elicitation). No question was repeated. At the beginning of each part, participants answered three trial questions to familiarize them with the experimental environment.

For both Time and Risk, the alternatives were divided into two groups: four MAIN alternatives, which are presented in Table 1 and Table 2, and some "confounding" alternatives, described in the Online Appendix. Each individual solved all 11 choice problems involving the MAIN alternatives.[25] The other questions were set in order to obtain particular information about rationality: monotonicity, impatience,[26] stochastic dominance, as well as about possible behavioural effects: choice overload, compromise effect, attraction effect. The positions of the alternatives were randomized. The subjects could face two orders of questions; we also inverted Time and Risk elicitation, resulting in a total of four treatments which are described in section 2 of the Online Appendix.

After the choice elicitation part, subjects were asked to rank the four MAIN alternatives. No indifferences were permitted; hence the reported welfare relation is always a linear order. Subsequently, subjects filled in a questionnaire about their comprehension of the experimental design and criteria of choice in both Time and Risk. The question-

[25]The collection of all non-empty subsets of the MAIN alternatives is a crucial feature of the design for two reasons: (1) it creates a symmetric portion of the dataset that allows for immediate comparisons across environments (Time and Risk), and (2) beyond the symmetry, which is also satisfied by binary sets, as in Agranov & Ortoleva (2017), MAIN sets allow us to infer welfare on sets that are more complex than binary sets but at the same time do not contain behavioural effects. This allows us to test the capacity of welfare methods such as **MS**, **TC** and **EIG** to break cycles in the **CRP** relation.

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

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.

naire is presented and analysed in Section 4 of the Online Appendix.

The average reward was about £ 19 per subject and the experiment lasted on average 75 minutes. The reward was measured in Tokens 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 the last screen informed subjects of their earnings in each part.

All sessions were conducted at the University of St. Andrews between June and September 2019. Undergraduate and postgraduate students were recruited voluntarily. Eleven sessions were run with a total of 145 subjects. No subject participated in more than one session. Their earnings were paid via bank account at the end of the experiment and at future dates, as specified by both the instructions and the experimenter. The experiment was completely anonymous and all subjects signed a consent form wherein they agreed to provide their UK bank account number and sort code. The experiment was performed using z-tree (Fischbacher, 2007).

3.1 Experimental Hypotheses

In the subsequent part of the paper, we test Informational Responsiveness in both its informational (Propositions 1 and 2) and frequency interpretations (Proposition 3). In view of our theoretical results, Figure 1 summarizes the properties of welfare methods and it is the reference point for our experimental hypotheses. We use the notation $\mathsf{Prop}(2)$ and $\mathsf{Prop}(3)$ to denote those welfare methods that are equivalent to **CRP** under the conditions of Propositions 2 and 3, respectively. Notice that, **MS** and **TC(CRP)** do not satisfy CON, hence they do not fall under the frequency interpretation of IR that we discussed in Section 2.1. However, our experimental data show that $P^D_{\mathbf{CRP}}$ is extremely well-behaved: 99.3% of subjects have acyclic $P^D_{\mathbf{CRP}}$ in Time, and 97.2% in Risk; while 93.8% have transitive $P^D_{\mathbf{CRP}}$ in Time, and 88.3% in Risk. This is, first, a testimony of the relevance of the class of models ($\mathcal{MON}$) studied in our theoretical section, and second, it assures an alignment between the performance of the welfare methods characterized by Proposition 3. Therefore, in interpreting our results, it is safe to generalize the frequency interpretation of IR to **MS** and **TC(CRP)**; in fact, these methods will overlap almost perfectly with **CRP**.

	NEU	CNN	IR	CON	Prop(2)	Prop(3)
CRP	√	√	√	√	√	√
MS	√	√	√	×	√	√
TC(CRP)	√	√	√	×	√	√
EIG	√	√	√	×	√	×
CC	√	√	√	×	√	×
SEQ	√	√	×	×	×	×
BR	√	√	×	√	×	×
TC(BR)	√	√	×	×	×	×
ATT(MS)	√	×	×	×	×	×
ATT(BR)	√	×	×	×	×	×

Figure 1: Properties of welfare methods.

Hypothesis 1. *A substantial proportion of the subjects violate WARP, therefore standard revealed preference analysis does not elicit a transitive and complete welfare relation.*

Hypothesis 1 is the premise of our theoretical and experimental exercise. If subjects behave consistently then there would be no need for behavioural welfare analysis. Hence, we start showing that, in our experiment as in many others in the literature, subjects repeatedly violate WARP.

Hypothesis 2. *Welfare methods that satisfy IR are more effective in eliciting welfare relations than the remaining welfare methods in both homogeneous and non-homogeneous domains.*

Hypothesis 2 tests the informational interpretation of IR as well as the extent to which the class of models $\mathcal{MON}$ can represent our experimental data.

Hypothesis 3. *Let IR be satisfied; (i) welfare methods that fall under Proposition 3 are more effective in eliciting welfare relations than the remaining welfare methods in non-homogeneous domains, (ii) but not in homogeneous domains.*

Hypothesis 3 tests the frequency interpretation of IR as well as the extent to which the class of models $\mathcal{APU}$ can represent our experimental data.

Hypothesis 4. *The* ***TC*** *and the* ***ATT*** *methods are more effective in eliciting welfare relations when based on* ***CRP*** *and* ***MS*** *respectively, than when based on* ***BR****.*

Hypothesis 4 aims to provide further evidence in favour of IR.

Hypothesis 5. *Welfare methods that satisfy IR are more effective in eliciting welfare relations more and more observations are added to the dataset.*

Hypothesis 5 deals with the monotonic interpretation of IR. If the choice of an alternative is welfare-relevant, then methods that satisfy IR should be more and more effective in eliciting the welfare relation as more and more data are added to the dataset.

Hypothesis 5 is founded on the strong assumption that subjects reveal their welfare regardless of the type of decision problem they face. The literature has provided countless examples that contradict this assumption. We verify which decision problems are welfare-relevant in our experiment and in doing so we provide a direct test of IR (Hypothesis 6).

Hypothesis 6. *Subjects, on average, reveal their welfare in every choice problem. Hence, IR is everywhere satisfied.*

4 Experimental Results

4.1 Premise: Do individuals consistently reveal welfare?

Figure 2 presents the distribution of WARP violations in Time, Risk, and in a random behaviour benchmark. For each subject *i*, WARP violations are determined as the number of cycles of length 2 in the graph of revealed preference.

$$WARP_i = \sum_{x,y} C_{xy} \cdot C_{yx}$$

Hypothesis 1 is confirmed. In Time, 63% of the subjects violate WARP at least once, while in Risk this percentage grows to 94%. More generally, subjects violate WARP less in Time than in Risk (the distributions are significantly different - two-sample Chi-Square test yields p-value=0.000). Furthermore, subjects do not behave randomly (p-value=0.000 for both Time and Risk). In summary, the data show a fundamental difference in the behaviour of the subjects in the two environments. This finding is confirmed, in Figure 3, by the low correlation between the number of WARP violations in Time and Risk. In view of this evidence, we will treat Time and Risk separately in the analysis of the welfare methods.

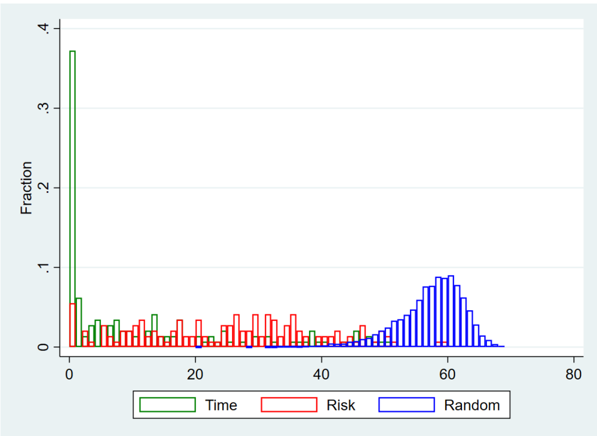

Figure 2: Distribution of the violations of WARP.

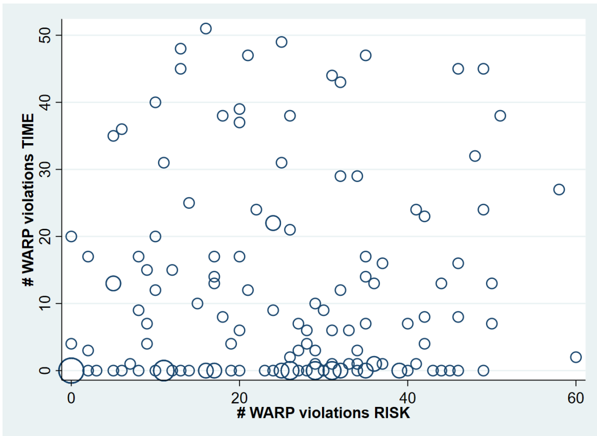

Figure 3: Violations of WARP in Time and Risk.

4.2 Identification of the reported welfare relation

In this subsection, we evaluate the welfare methods by matching their elicited welfare relation with the reported ones in both Time and Risk, using the ALL dataset (non-homogeneous domain), MAIN sets and BINARY sets (homogeneous domains). Our identification exercise is twofold. First, we focus on identifying the reported best element. We use a unique and set identification exercise; namely, we identify when the reported best element is in the set of maximal elements of the elicited welfare relation. The difference between the proportion of subjects that a welfare method uniquely and set identifies will give a simple measure of its coarseness. We complement these exercises by assuming that a risk-neutral policymaker picks from the set of maximal elements endowed with a uniform distribution, and, in doing so, we provide a measure of expected identification. Second, we focus on identifying the entire welfare relation. Again, we do so both in a unique and set identification exercise. In this case, to provide a finer measure of distance between the identified and reported welfare relation, we use two measures of similarity: the symmetric difference and the reverse asymmetry measure.

Let M 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 $\mathsf{REP}_i(\succ)$. Our measures of identification are the following:

- Unique Identification [**UI**]:
$$\frac{\#\{i \in M : \max[\mathsf{REP}_i(\succ)] = \max[f_i(D)]\}}{\#M}$$
A measure of set identification of the best element [**SI**] is obtained by substituting $\max[\mathsf{REP}_i(\succ)] = \max[f_i(D)]$ with $\max[\mathsf{REP}_i(\succ)] \subseteq \max[f_i(D)]$.

- Expected Identification [**EI**]:
$$\frac{\sum_{i \in M : \max[\mathsf{REP}_i(\succ)] \in \max[f_i(D)]} \frac{1}{\#\{\max[f_i(D)]\}}}{\#M}$$

- Welfare Relation Identification [**WRI**]:
$$\frac{\#\{i \in M : \mathsf{REP}_i(\succ) = f_i(D)\}}{\#M}$$
Similarly as above, the measure of set identification [**SWRI**] is obtained by substituting $\mathsf{REP}_i(\succ) = f_i(D)$ with $\mathsf{REP}_i(\succ) \subseteq f_i(D)$.

- The Symmetric Difference [**SD**] between $f_i(D)$ and $REP_i(\succ)$. The symmetric difference $\triangle$ between two binary relations R_1, R_2 is defined as follows:
$$R_1 \triangle R_2 = (R_1 \setminus R_2) \cup (R_2 \setminus R_1)$$

- The Reverse Asymmetry measure [**RA**], denoted here as $\triangledown$, between $REP_i(\succ)$ and $f_i(D)$. It is defined as the number of times the asymmetric part of the reported order is reversed - namely, given two asymmetric binary relations P_1 and P_2:
$$P_1 \triangledown P_2 = \{(x, y) \in P_1 : (y, x) \in P_2\}$$

The rationale behind the use of two measures of similarity is simple. On one hand, SD equally considers the symmetric and asymmetric part of the binary relation, hence penalizing coarse methods such as **BR**. On the other hand, RA focuses only on the asymmetric parts, disentangling those differences that are in principle worse and so penalizing methods that are more likely to map - or that can only map - into asymmetric binary relations, such as **EIG** and **SEQ**.

4.2.1 Time

The results in the following subsections are reported in relation to the previously stated hypotheses.

Hypothesis 2 - Table 3 shows that methods that satisfy IR perform significantly better than **BR** both uniquely (32-33%) and in expectation (13-14%). This result generalizes to any other method that does not satisfy IR (except for **SEQ**, see below). Importantly, since **BR** identifies only those subjects that rationally reveal their best element, the 32-33% gap relates to subjects who violated WARP in choosing their best element, but these latter were nonetheless identified correctly. Focusing on WRI, the left part of Table 4 also shows that methods that satisfy IR perform better than **BR** by 10-15%. Further, in both Tables, the set identification is very close to the one achieved by **BR** which is clearly an upper bound.[27]

Table 4 contains further relevant information when we look at the remaining methods that violate IR. For instance, we notice that **SEQ** performs in line with **MS** and **TC(CRP)** and better than **EIG**. To interpret this piece of evidence, we look at SD and RA in the right part of Table 4.[28] We can see that **SEQ** and **EIG** are significantly outperformed by **MS** and **TC(CRP)** in terms of SD and RA. In words, the linearly ordered structure of **SEQ** and **EIG** may allow a better "point identification" of the welfare relation but, when this is not identified, the mistakes appear to be substantial.

Hypothesis 3 - To test hypothesis 3, we look at the top parts of both Tables 3 and 4. We can see that **CC** is outperformed in terms of UI, EI, and WRI but only in the non-homogeneous domain (ALL). **EIG** is also outperformed in terms of WRI. This result

[27]The set identification exercise shows a very small misalignment between elicited and reported welfare relation, in particular regarding the best element, as **BR** set identifies 94% of the subjects and the welfare methods that satisfy IR & RP set identify close 90% of the subjects.

[28]Notice that, as theoretically predicted, **BR** induces a lower bound on RA due to its cautious approach and, with the exception of **TC(BR)**, an upper bound on SD due to its coarseness.

confirms the importance of both the informational and frequency interpretations of IR as well as providing suggesting evidence that, in our experiment, part of the subjects behave according to the models characterized by Proposition 3 [$\mathcal{MON}$], but not by those characterized by Proposition 2 [$\mathcal{APU}$].

Hypothesis 4 - We turn our attention to the methods **TC** and **ATT**. It is clear in both tables that when these methods are based on **CRP** and **MS** respectively, they identify substantially more subjects providing more evidence about the importance of IR.

Hypothesis 5 - In Table 3, the power of identification for methods that satisfy IR is increasing in the size of the dataset, suggesting that individuals reveal information about welfare throughout all dataset. In Table 4, this result is less straightforward, providing interesting insights. If we observe the SD of methods that satisfy IR, we notice that it is decreasing for all methods apart from **EIG** and **CC**. This observation confirms the importance of the frequency interpretation of IR and the generalization from the class of models $\mathcal{APU}$ to $\mathcal{MON}$. Based on the representation proposed in Proposition 3 (**WCRP**$^+$), we also conjecture that, in Time, Binary sets are particularly relevant to elicit welfare relations, while some sets with potential behavioural effects are not. In section 4.5.1 we provide suggestive evidence for both conjectures.

Table 3: UNIQUE & EXPECTED IDENTIFICATION - TIME

	METHODS	UI (SI) ALL	UI (SI) MAIN	UI (SI) BINARY	EI ALL	EI MAIN	EI BINARY
Prop(3)	CRP	0.87 (0.89)	0.81 (0.88)	0.77 (0.77)	0.88	0.84	0.77
	MS	0.87 (0.89)	0.81 (0.89)	0.79 (0.83)	0.88	0.85	0.80
	TC(CRP)	0.87 (0.89)	0.81 (0.88)	0.77 (0.88)	0.88	0.84	0.81
Prop(2)	EIG	0.87 (0.87)	0.83 (0.85)	0.79 (0.83)	0.87	0.84	0.80
	CC	0.81 (0.86)	0.83 (0.88)	0.77 (0.86)	0.84	0.86	0.81
No-IR	ATT(MS)	-	0.80 (0.86)	-	-	0.83	-
	ATT(BR)	-	0.65 (0.90)	-	-	0.77	-
	TC(BR)	0.66 (0.89)	0.69 (0.90)	0.77 (0.88)	0.76	0.79	0.81
	BR	0.55 (0.94)	0.67 (0.93)	0.77 (0.77)	0.74	0.79	0.77
	SEQ	-	0.83 (0.83)	-	-	0.83	-

NOTES -- On the left, we show the proportion of subjects for whom each method uniquely and set identifies the reported best element. On the right, is the expected proportion of subjects for whom each method identifies the reported best element. When methods "set identify" the best element - namely, they identify more than one element as best, we pick uniformly from the set of identified best elements.

Table 4: IDEN. WELFARE RELATION, SD & RA - TIME

	METHODS	WRI (SWRI) ALL	WRI (SWRI) MAIN	WRI (SWRI) BINARY	ALL SD	ALL RA	MAIN SD	MAIN RA	BINARY SD	BINARY RA
Prop(3)	CRP	0.61 (0.68)	0.57 (0.68)	0.59 (0.59)	2.14	0.93	2.27	0.87	2.62	1.31
	MS	0.62 (0.68)	0.59 (0.69)	0.61 (0.68)	2.17	0.98	2.24	0.90	2.60	1.05
	TC(CRP)	0.61 (0.68)	0.58 (0.70)	0.59 (0.73)	2.14	0.87	2.24	0.81	2.79	0.85
Prop(2)	EIG	0.54 (0.54)	0.60 (0.61)	0.60 (0.68)	2.64	1.31	2.44	1.20	2.64	1.07
	CC	0.54 (0.63)	0.58 (0.70)	0.59 (0.72)	2.55	1.08	2.21	0.88	2.60	0.93
No-IR	ATT(MS)	-	0.59 (0.68)	-	-	-	2.35	0.96	-	-
	ATT(BR)	-	0.50 (0.74)	-	-	-	2.89	0.70	-	-
	TC(BR)	0.44 (0.69)	0.52 (0.73)	0.59 (0.73)	3.19	0.85	2.73	0.76	2.79	0.85
	BR	0.42 (0.74)	0.50 (0.72)	0.59 (0.59)	3.14	0.54	2.69	0.64	2.62	1.31
	SEQ	-	0.60 (0.60)	-	-	-	2.31	1.15	-	-

NOTES -- On the left, we show the proportion of subjects for whom each method uniquely and set identifies the entire reported welfare relation. On the right, "SD" and "RA" denote respectively symmetric difference and reverse asymmetry. We report the mean of SD and RA over all subjects who have violated WARP at least once over the ALL dataset. To interpret these measures, note that SD can take values between 0 and 12, while RA can take values between 0 and 6.

4.2.2 Risk

Hypothesis 2 - Table 5 shows that methods that satisfy IR perform significantly better than **BR**, both uniquely (47-49%) and in expectation (18-20%). Similarly, the left part of Table 6 confirms this finding **BR** is outperformed in the WRI exercise by 15-20%. Again, the result can be generalized to all the remaining methods that violate IR. The notable exception is again **SEQ**. Using the same reasoning previously adopted, we observe that, in the right part of Table 6, **SEQ** performs significantly worse than **MS** and **TC** in terms of both SD and RA.

Interestingly, we notice that the high number of violations of WARP have induced a lower alignment between elicited and reported welfare relation and an extremely coarse welfare relation elicited by **BR** as we can see comparing its unique (12%), set (88%), and expected (43%) identification results. Nonetheless, applying the transitive core principle on **BR** (Nishimura, 2018), we can see that the set identification result (68%) drops close to the level of the welfare methods that satisfy IR (60%-65%). This observation confirms that the potential elicitation of methods that satisfy IR is not dissimilar from the much coarser **TC(BR)**.

Hypothesis 3 - Looking at the top part of Tables 5 and 6, we see that the **CC** method is generally outperformed by methods that fall under the class of Proposition 3. The results are not replicated for **EIG**. The reason has to be searched in lower alignment between the dataset and the models in Risk. We discuss this point further when analysing hypothesis 5 below.

Hypothesis 4 - As in Time, **ATT(BR)** and **TC(BR)** are outperformed by **ATT(MS)** and **TC(CRP)** providing more evidence about the importance of IR.

Hypothesis 5 - The power of identification is generally increasing in the size of the dataset with a steeper gradient than in Time. For example, focusing on **CRP**, we observe a greater gap between BINARY and MAIN sets and between MAIN and ALL sets in both Table 5 and 6. We also observe that Hypothesis 5 is confirmed by **EIG**. This observation allows us to discuss the alignment assumption which, in Risk, is weakened by the high number of violations of WARP. The **EIG** method is generally more sensitive to new observations than the other methods that satisfy IR. Hence, higher noise in the data does not compromise the capacity of **EIG** to elicit an asymmetric welfare relation. This result can be inferred comparing the point and set identification results of **EIG** with those of the other methods that satisfy IR in Tables 5 and 6. Furthermore, this observation suggests, on one hand, that sets different from the MAIN sets add valuable welfare information in Risk and, on the other hand, that BINARY sets are not as important as they are in Time to elicit welfare relations. Again, we provide suggestive evidence in favour of both these conjectures in section 4.5.2.

Table 5: UNIQUE & EXPECTED IDENTIFICATION - RISK

	METHODS	UI (SI) ALL	UI (SI) MAIN	UI (SI) BINARY	EI ALL	EI MAIN	EI BINARY
Prop(3)	CRP	0.59 (0.64)	0.52 (0.66)	0.42 (0.42)	0.61	0.59	0.42
	MS	0.59 (0.64)	0.52 (0.68)	0.46 (0.57)	0.61	0.60	0.50
	TC(CRP)	0.61 (0.65)	0.51 (0.70)	0.42 (0.66)	0.63	0.60	0.50
Prop(2)	EIG	0.61 (0.62)	0.60 (0.62)	0.46 (0.57)	0.62	0.61	0.50
	CC	0.55 (0.60)	0.56 (0.66)	0.42 (0.62)	0.57	0.61	0.50
No-IR	ATT(MS)	-	0.50 (0.59)	-	-	0.55	-
	ATT(BR)	-	0.26 (0.69)	-	-	0.49	-
	TC(BR)	0.28 (0.68)	0.33 (0.71)	0.42 (0.66)	0.44	0.50	0.50
	BR	0.12 (0.88)	0.17 (0.81)	0.42 (0.42)	0.43	0.49	0.42
	SEQ	-	0.55 (0.55)	-	-	0.55	-

NOTES -- On the left, we show the proportion of subjects for whom each method uniquely and set identifies the reported best element. On the right, is the expected proportion of subjects for whom each method identifies the reported best element. When methods "set identify" the best element - namely, they identify more than one element as best, we pick uniformly from the set of identified best elements.

Table 6: IDEN. WELFARE RELATION, SD & RA - RISK

	METHODS	WRI (SWRI) ALL	WRI (SWRI) MAIN	WRI (SWRI) BINARY	ALL SD	ALL RA	MAIN SD	MAIN RA	BINARY SD	BINARY RA
Prop(3)	CRP	0.24 (0.37)	0.19 (0.37)	0.20 (0.20)	3.21	1.37	3.35	1.24	4.09	2.04
	MS	0.24 (0.38)	0.20 (0.39)	0.21 (0.30)	3.24	1.40	3.32	1.32	4.18	1.77
	TC(CRP)	0.24 (0.37)	0.19 (0.42)	0.20 (0.37)	3.19	1.34	3.28	1.15	4.19	1.35
Prop(2)	EIG	0.27 (0.31)	0.26 (0.28)	0.20 (0.30)	3.29	1.60	3.30	1.60	4.26	1.81
	CC	0.21 (0.28)	0.19 (0.37)	0.20 (0.32)	3.33	1.47	3.32	1.36	4.18	1.60
No-IR	ATT(MS)	-	0.22 (0.35)	-	-	-	3.48	1.45	-	-
	ATT(BR)	-	0.13 (0.44)	-	-	-	3.79	1.16	-	-
	TC(BR)	0.10 (0.40)	0.12 (0.40)	0.20 (0.37)	4.00	1.18	3.81	1.18	4.19	1.35
	BR	0.06 (0.63)	0.10 (0.54)	0.20 (0.20)	4.35	0.63	4.01	0.85	4.09	2.04
	SEQ	-	0.25 (0.25)	-	-	-	3.51	1.76	-	-

NOTES -- On the left, we show the proportion of subjects for whom each method uniquely and set identifies the entire reported welfare relation. On the right, "SD" and "RA" denote respectively symmetric difference and reverse asymmetry. We report the mean of SD and RA over all subjects who have violated WARP at least once over the ALL dataset. To interpret these measures, note that SD can take values between 0 and 12, while RA can take values between 0 and 6.

4.3 Optimal Weighting

So far, our evidence in favour of IR is based on the comparison between welfare methods, hence it is both indirect and relative. We aim to generalize our results in both directions.

First, we provide an absolute measure of performance that takes into account the misalignment between reported and elicited welfare relations. In principle, we have partially answered this question by providing both unique and set identification results. Now, we approach this problem more structurally, comparing each welfare method with a data-driven benchmark that we call Optimal Weighting method [**OW**].

Second, since, as shown below, **OW** belongs to the class of methods **WCRP** characterized in Proposition 3, it provides us with a direct test of IR. If **OW** optimally assigns strictly positive weights to all observations, then IR is satisfied (Hypothesis 6)?

To define **OW**, we divide the entire dataset into five parts: binary sets [B], ternary sets [T], a quaternary set [Q], sets with asymmetric dominance [AD], and 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 vector of weights is $\mathbf{w} = (w_B, w_T, w_Q, w_{AD}, w_{BIG})$. We define the method **OW** as follows:

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

where $OW_{xy} = \sum_{i \in \Gamma} w_i C^i_{xy}$ and $\Gamma = \{B, T, Q, AD, BIG\}$.

Weights are calculated by optimizing the sum of two measures: (1) expected identification of the maximal element [EI] and (2) unique identification of the entire welfare relation [WRI]. We recall that 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 identifies the entire reported welfare relation. The optimization problem is as follows:[29]

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

[29]Note that weights attached to different parts of the dataset may be negative. Consider the following example on MAIN sets: a subject always chooses **x** when available, **y** when **x** is not available, and **z** from { **z**,**w** }. Then, he reports **x** ≻ **y** ≻ **w** ≻ **z**. In this case, since **x**, **y** are clearly best, binary sets receive a small negative weight that guarantees **w** ≻ **z**, and does not change the other preferences that are revealed by sets with three and four elements.

where for each subject i:

$$xR^D_{f_i}y \quad \Leftrightarrow \quad \mathbf{w} \cdot \mathbf{C}_{xy_i} \geq \mathbf{w} \cdot \mathbf{C}_{yx_i}$$

4.4 Completeness of the methods

In this section, we compare the identification results of welfare methods with the data-driven method **OW** and refer to the distance between them as the completeness of the methods. We borrow the term "completeness" from Fudenberg et al. (2022). In their paper, the authors use machine learning to measure the amount of variation in the data that a theory can capture. Their notion 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. (2022). In our framework, we define completeness, denoted as $\texttt{Com}(f)$ for a welfare method f, as follows:

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

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 $f_L = \textbf{BR}$ and $f_U = \textbf{OW}$, since the former identifies only perfectly rational subjects while the latter is based on the knowledge of the reported welfare relation which is generally not available to the researcher. Table 7 shows the completeness of the methods across different types of identification procedures in both Time and Risk.

Table 7: COMPLETENESS OF THE METHODS

	TIME			RISK		
METHODS	UI	EI	WRI	UI	EI	WRI
CRP	0.92	0.89	0.79	0.87	0.80	0.67
MS	0.92	0.89	0.85	0.87	0.80	0.67
TC(CRP)	0.94	0.91	0.82	0.91	0.86	0.67
EIG	0.92	0.82	0.53	0.92	0.81	0.77
CC	0.76	0.62	0.50	0.81	0.62	0.54
ATT(MS)	0.72	0.59	0.71	0.72	0.51	0.59
ATT(BR)	0.28	0.14	0.35	0.27	0.25	0.26
TC(BR)	0.30	0.40	0.09	0.29	0.63	0.15
SEQ	0.82	0.59	0.76	0.81	0.53	0.69
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 (UI), expected (EI) and entire (WRI) identification procedures.

Since **BR** and **OW** are respectively the lower and upper bound for our identification analysis, they respectively take the values of zero and one. Methods that satisfy IR and RP have generally higher completeness than other methods. Note that in both UI, EI, and WRI, and both in Time and Risk, there is always at least one method that satisfies both IR and RP that is more complete than every method that fails to satisfy them.

4.5 Direct test of Informational Responsiveness

Finally, we propose a direct test for IR (Hypothesis 6). We focus on the family of methods $\mathbf{WCRP}^{+}$. As mentioned in section 4.3, if each choice receives a strictly positive weight independently from the set where it happened, then IR is satisfied.

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. Tables 8 and 9 report results based on six objective functions. Before presenting and discussing them, two clarifications are needed. First, the optimization problem described in section 4.3 may have non-unique results. In such cases, we report the minimum and maximum weights for each part of the dataset such that there exists a system of weights that solves the optimization problem. Importantly, this does imply that any vector of weights in the cartesian product of the intervals guarantees optimal identification. Second, if choices from a particular part of the dataset are irrelevant then this part receives a positive, negative or zero weight without changing the result.

4.5.1 Time

In Table 8, we observe that strictly positive weights are associated with 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),[30] while they have negative weights when we identify the entire welfare relation. The former result is expected; the latter is somewhat surprising since it shows that subjects wrongly reveal their welfare in this part of the dataset.

Table 8: OPTIMAL WEIGHTS - TIME

	TIME				
IDENTIFICATIONS	**BIN**	**TER**	**QUA**	**BIG**	**AD**
UI	[0.6,1]	[0.2,0.4]	[0.1,0.2]	[0.6,1]	[-0.2,1]
EI	[0.6,1]	[0.2,0.4]	[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	[0.6,0.7]	[0.6,0.7]	[0.6,0.7]	[0.6,0.7]	-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 in the paper to define OW.

We also find that binary sets are particularly important throughout all the possible objective functions. This explains both the relatively good performance of methods on these sets (Table 3) and the fact that the identification power of **EIG** and **CC** decreases in the size of the sets, as observed in Table 4.

4.5.2 Risk

Table 9 shows that, in Risk, IR binds everywhere, since strictly positive weights are attached to any domain. There are two exceptions. Firstly, AD sets are irrelevant when we

[30]In Tables 8 and Table 9, AD sets receive both positive and negative weights ([-0.2,1] in Time and [-0.2,0.8] in Risk) when the objective functions are either UI or EI. This is due to the fact that only two alternatives out of the four MAIN alternatives are represented in AD sets and often these alternatives are not reported and chosen as the best alternatives. Therefore, from this interval we cannot conclude anything about the importance of AD sets and to analyse them we need to focus on weights assigned to AD sets under the remaining four objective functions.

focus only on the reported best element, but this observation does not convey relevant information as previously mentioned. On the contrary, AD sets receive strictly positive weights, differently from Time, in the other four objective functions. This shows that, in Risk, behavioural effects such as attraction and compromise effect do not seem to undermine the elicitation of preferences. This result is perfectly in line with our analysis in section 4.2.2.

Table 9: 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]	[-0.2,0.8]
EI	[-0.2,-0.1]	[0.3,0.8]	[0.5,1]	[0.4,0.9]	[-0.2,0.8]
WRI	[0.3,0.7]	[0.5,1]	[0.8,1]	[0.4,0.8]	[0.4,1]
SD	0.4	0.4	1	0.3	0.4
SD & RA	[0.2,0.4]	[0.2,0.4]	[0.4,0.8]	[0.2,0.4]	[0.2,0.4]
EI & WRI	[0.1,0.3]	[0.3,0.5]	[0.6,1]	[0.4,0.6]	[0.4,0.6]

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 in the paper to define OW.

Second, 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 the previous sections. In Table 5 we find that welfare methods perform poorly on binary sets and, in Table 6, that the **EIG** method has an increasing identification power in the size of the sets. 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.[31]

5 Conclusion

In this paper, we study behavioural welfare analysis using an approach in line with Bernheim & Rangel (2009). We claim that model-less (or model-free) approaches do not exist. We introduce a novel terminology that distinguishes between "implicitly model-based" (Bernheim & Rangel (2009), Apesteguia & Ballester (2015), Nishimura (2018),

[31]This evidence suggests further research on the role of attention in choices among gambles and it is in line with stochastic models such as Manzini & Mariotti (2014a) and Cattaneo et al. (2020).

Horan & Sprumont (2016)) and "explicitly model-based" approaches (Masatlioglu et al. (2012), Rubinstein & Salant (2012), Manzini & Mariotti (2014b)). We axiomatically analyse revealed preference mappings: from choices to welfare preference relations. We propose an appealing property, called Informational Responsiveness, that aims to solve the informational drawbacks of some of the approaches proposed by the literature. In a series of theoretical results, we show that Informational Responsiveness is an important and desirable condition for behavioural welfare analysis as it avoids paradoxical welfare conclusions and implicitly selects a broad family of models, that we denote preference-monotonic models.

Using a novel experimental design, we test Informational Responsiveness in both its premise and its conclusion. First, we show that individuals repeatedly violate the Weak Axiom of Revealed Preference both in time and risk preferences. This calls for the adoption of behavioural (non-standard) welfare analysis. Second, we find that welfare methods that satisfy Informational Responsiveness 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. (2022). Finally, combining our theoretical results with the use of an optimal weighting algorithm, we directly test Informational Responsiveness. We show that, in our experiment, subjects reveal welfare in all parts of the dataset confirming the role of Informational Reposnsiveness.

A Proofs

A.1 Proposition 1

In the following proofs, we omit the subscript f to ease the reading. By the alignment assumption, we have $x \geq y$ if and only if $C_x \geq C_y$ on datasets $\mathcal{A}$. Hence, we have to prove that **CC** is fully characterized by IR, CNN, and NEU. It is trivial to show that **CC** satisfies all three axioms, hence we prove the converse.

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 $xI^D y$. Suppose to the contrary that $xP^D y$ 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 $yP^D x$, 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 $xP^D y$. If $C_x + C_y = 2$ and $C_x > C_y$ then $xP^D y$ by CNN; if $C_x = C_y$ then $xI^D y$ 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 $xP^D y$ by IR and the inductive hypothesis; if $C_x - C_y > 1$ then $xP^D y$ by CNN and the inductive hypothesis; if $C_x = C_y$ then $xI^D y$ by NEU.

A.2 Proposition 2

By Theorem 1 in Fudenberg et al. (2015), in $\mathcal{APU}$ models, probability ratios satisfy an ordinal version of IIA, namely, they can be rescaled by a strictly monotonic function and they are equally rescaled in every menu. Therefore, there exists a preference relation $\succeq$ that ranks the alternatives preserving their frequency of choices within and between menus. $\mathcal{APU}$ models guarantee that $x \succeq y$ if and only if $xR^{\mathtt{hom}(D)}_{\mathbf{CRP}} y$ (it is trivial to check that this result does not hold on non-homogeneous domains). Hence, we need to prove that **CRP** is characterized by CNN, IR, NEU and transitivity on $\mathtt{hom}(D)$. **CRP** trivially satisfy CNN, IR, and NEU, but it is generally non-transitive. However, on can see that, under $\mathcal{APU}$ models, **CRP** is transitive on $\mathtt{hom}(D)$, where it is equivalent to **CC**.

We start by stating two related observations. First, by finiteness of X, transitivity and completeness of R^D, there exists a real-valued function u on X such that for all $x, y \in X$; $xR^D y$ if and only if $u(x) \geq u(y)$. Second, let $\phi : X \to R^{n-1}$, where $|X| = n$, be a vector-valued function that represents R^D and $\phi(x)_z$ be the valued assigned to x when compared to z. By transitivity, ϕ and u are aligned. If not, we may observe $\phi(x)_y > \phi(y)_x$, $\phi(y)_z > \phi(z)_y$ and $\phi(z)_x \geq \phi(x)_z$. In words, if a welfare method assigns a value to x that changes non-monotonically with u with the compared alternatives, for every $\mathcal{APU}$ model, we can construct a set of alternatives X and a homogeneous domain such that transitivity is violated. **CRP** is the only (generally) non-transitive method that is always transitive under $\mathcal{APU}$ models on homogeneous domains and satisfies IR, NEU and CNN.

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}; $E = E_{xy} = E_{yx} = \sum_{z \neq x,y} \sum_{x,y \notin A} D(z,A)$.

Assume w.l.o.g. that $u(x) \geq u(y)$. NEU implies $xI^D y$ on B and E. By the reasoning in Proposition 1, $C_{x,-y} \geq C_{y,-x}$ and $C_{xy} \geq C_{yx}$ imply $xR^D y$. Note that, $C_{x,-y} \geq C_{y,-x}$ and $C_{xy} \geq C_{yx}$ only hold on $\texttt{hom}(D)$.

To complete the proof we need to extend the argument to D_{xy} and D_{yx}. 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 attaches a positive value to D_{xy}, D_{yx} could lead to $yR^D x$ when $u(x) > u(y)$. Transitivity rules out this possibility. In fact, 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. So, suppose by contradiction that $u(x) > u(y)$ and $yR^D x$; it must be that the value attached to observations in D_x, D_y is positive, since $D_y > D_x$. We proved that $xP^D y$ over the parts of the dataset denoted as $C_{x,-y}$, $C_{y,-x}$,C_{xy}, C_{yx}, B, E. Suppose we add an observation (x,A) with $y \in A$. D_y increases by a positive value and since we assumed $yR^D x$, CNN is violated.

Independence of transitivity For all $x, y \in X$ and $D \in \texttt{hom}(D)$:

$$N_{xy} \geq N_{yx} \Leftrightarrow xR^D y$$

where $N_{xy} = C_{xy} + \delta \cdot D_{xy}$ with $\delta > 0$. Take a Luce model where $u(x) = 3$, $u(y) = 2$ and $u(z) = 1$. $p(x,A) = \frac{u(x)}{\sum_{y \in A} u(y)}$. Assume we have 1000 observations for each binary set. $N_{xy} = 600 + \delta \cdot 250$, and $N_{yx} = 400 + \delta \cdot 333$. For $\delta > \frac{200}{83}$, $yP^D x$. Note that, when the number of alternatives increases, D_{xy} increases. Hence, a similar contradiction can be found for any small δ. The above method satisfies NEU, IR and CNN, but for $\delta = 2$, it yields the cycle $xP^D yP^D zP^D x$ violating transitivity.

The result does not hold on $\mathcal{MON}$ Take the following dataset on $\texttt{hom}(D)$ with 1000 observations for each binary set. $D(x, \{x,y\}) = 900$, $D(y, \{y,z\}) = 600$, $D(x, \{x,z\}) = 600$. This dataset can be the outcome of a model in $\mathcal{MON}$ with $x \geq y \geq z$. However, $C_y = 700$ and $C_z = 800$. Notice also that $zP_{\mathbf{EIG}}y$ which proves that **EIG**, which does not satisfy CON, may not be equivalent to **CRP** under models $\mathcal{MON}$.

A.3 Proposition 3

Similarly to Propositions 1 and 2, we first we notice that $C_{xy} \geq C_{yx}$ if and only if $x \geq y$ under the set of models $\mathcal{MON}$. These models together with our alignment assumption imply that for all $A \subseteq X$, $C^A_{xy} \geq C^A_{yx}$. Further, they imply that $C_{xy} \geq C_{yx}$ if and only if $C^A_{xy} \geq C^A_{xy}$ for all $A \subseteq X$ if $x, y \in A$. Hence, $xR_{\mathbf{CRP}}y$ if and only if $xR_{\mathbf{WCRP}^+}y$. We can therefore show that the method **CRP** is fully characterized by IR, NEU, CNN, and CON. One direction of the proof is immediate since **CRP** satisfies all these requirements. Conversely, the proof follows the arguments above.

We exploit the partition of the dataset into C_{xy}, C_{yx}, $C_{x,-y}$,, $C_{y,-x}$, B, D_{xy}, D_{yx}, and E. Assuming $x \geq y$, NEU and CON imply xI^Dy on B and E, and now also on $C_{x,-y}$,, $C_{y,-x}$, D_{xy}, and D_{yx}. Hence, the welfare relation is only driven by C_{xy} and C_{yx}. Since $x \geq y$ if and only if $C_{xy} \geq C_{yx}$; IR, NEU, and CNN imply xR^Dy.

A.4 Proposition 4

Notice the following trivial observation:

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

Hence, the number of swaps can be rewritten as:

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

Define a new measure $\Delta(C,P)$ that, equivalently to the swaps distance, defines the degree of similarity between a dataset and a linear 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

$$d_s(C,P_1^*) \leq d_s(C,P_2^*) \Leftrightarrow \Delta(C,P_1^*) \geq \Delta(C,P_2^*)$$

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 $d_s(C,P)$ increases by $n \in \mathcal{N}$, 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 $C_{xy} \geq C_{yx}$ 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 $P_{\mathbf{MS}}$.

B Weak Informational Responsiveness

The reader may be at odds with the idea that a single observation should have the power to modify a judgment of indifference. If this is the case, the underlying relevance of the asymmetric part of the welfare relation seems to be infinitely more important than its symmetric part. A simple weakening of IR allows us to retain its informational interpretation and introduce some flexibility in its monotonic interpretation.

Let $(A_i)_{i=1}^I$ be a collection of sets, and interpret $(x,(A_i)_{i=1}^I)$ as x is chosen from all the sets in the collection, and $y \in (A_i)_{i=1}^I$ as y being an element in all these sets.

Axiom 6 (Weak Informational Responsiveness **[WIR(I)]**).

$$xI_f^D y \quad \& \quad x,y \in (A_i)_{i=1}^I \quad \Rightarrow \quad xP_f^{D+(x,(A_i)_{i=1}^I)}y$$

As for the main version of IR, WIR(I) has to be completed with the consequent that deals with removing observations. Hence, the following consequent: if a collection of observations $(x,(A_i)_{i=1}^I)$ exists in the data, then $yP_f^{D-(x,(A_i)_{i=1}^I)}x$.

We define the weak counting choice method **WCC(I)** for some strictly positive integer I as:

$$xP^D_{\mathbf{WCC(I)}}y \text{ if and only if } C_x - C_y > I$$

To complete this method with its symmetric part, $xI^D_{\mathbf{WCC(I)}}y$ if and only if $\neg xP^D_{\mathbf{WCC(I)}}y$.

Proposition 5. *Given a collection of observations on a dataset $\mathcal{A}$, a method g satisfies WIR(I), NEU and CNN if and only if $g =$ **WCC(i)** with $1 \leq i < I$.*

Proof. The argument is only a slight modification of Proposition 1. NEU guarantees xI^Dy on symmetric collections of observations. Let $C_x + C_y = 1$ and $C_x > 0$. Then, either xI^Dy or xP^Dy. Suppose on the contrary that yP^Dx then CNN is violated. The same argument holds for any combination of C_x, C_y such that $C_x - C_y < I$ since NEU guarantees that xI^Dy if $C_x = C_y$ and CNN guarantees $\neg yP^Dx$ if $C_y > C_x$. Let $C_x - C_y = I$ then there exists a collection $(A_i)_{i=1}^I$ with x chosen from each set and y available, and this collection can be added to a symmetric dataset where $C_x = C_y$. Here, by NEU xI^Dy and by WIR $xP^{D+(x,(A_i)_{i=1}^I)}y$. Suppose by contradiction that xI^Dy; since a collection $(x, (A_i)_{i=1}^I)$ exists, by WIR $yP^{D-(x,(A_i)_{i=1}^I)}x$. However, by NEU $xI^{D-(x,(A_i)_{i=1}^I)}y$, hence a contradiction. Finally, let $C_x - C_y > I$. By CNN and NEU $\neg yP^Dx$. Suppose xI^Dy. Then, since there exists a collection $(x, (A_i)_{i=1}^I)$, by WIR $yP^{D-(x,(A_i)_{i=1}^I)}x$ which is a contradiction since $C_x > C_y$ implies either xI^Dy or xP^Dy. □

References

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

Aleskerov, F., Bouyssou, D., & Monjardet, B. (2007). Utility maximization, choice and preference.

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. (2016). Ordinal relative satisficing behavior: theory and experiments. *Working paper at SSRN: 2497947*.

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.

Bernheim, B. D., & Taubinsky, D. (2018). Behavioural public economics, chapter 5 in. *Handbook of Behavioral Economics - Foundations and Applications 1*, (pp. 318–516).

Bouacida, E., & Martin, D. (2021). Predictive power in behavioral welfare economics. *Journal of the European Economic Association*, *19*(3), 1556–1591.

Brady, R. L., & Rehbeck, J. (2016). Menu-dependent stochastic feasibility. *Econometrica*, *84*(3), 1203–1223.

Caplin, A., Dean, M., & Leahy, J. (2019). Rational inattentino, optimal consideration sets, and stochastic choice. *The Review of Economic Studies*, *86*(3), 1061–1094.

Cattaneo, M. D., Ma, X., Masatlioglu, Y., & Suleymanov, E. (2020). A random attention model. *Journal of Political Economy*, *128*(7), 2796–2836.

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.

Chetty, R., Looney, A., & Kroft, K. (2009). Salience and taxation: theory and evidence. *American Economic Review*, *99*(4), 1145–1177.

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

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

Enke, B., Gneezy, U., Hall, B., Martin, D. C., Nelidov, V., Offerman, T., & can de Ven, J. (2021). Cognitive biases: mistakes or missing stakes? *NBER Working paper 28650*.

Enke, B., & Zimmermann, F. (2019). Correlation neglect in belief formation. *Review of Economic Studies*, (86), 313–332.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, (10), 171–178.

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. (2014). Stochastic choice and revealed perturbed utility. *Report*.

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. (2022). Measuring the completeness of economic models. *Journal of Political Economy*, *4*(130), 956–990.

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. *Harvard Institute of Economic Research - Discussion Paper 2144*.

Hackethal, A., Kirchler, M., Laudenbach, C., Razen, M., & Weber, A. (2022). On the role of monetary incentives in risk preference elicitation experiments. *Journal of Risk and Uncertainty*, (p. Forthcoming).

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.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

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

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.

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

Luce, D. R. (1959). Individual choice behavior: a theoretical analysis.

Mandler, M., Manzini, P., & Mariotti, M. (2012). One million answers to twenty questions: choosing by checklists. *Journal of Economic Theory*, (147), 71–92.

Manzini, P., & Mariotti, M. (2007). Sequentially rationalizable choice. *American Economic Review*, *97*(5), 1824–1839.

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.

Manzini, P., Mariotti, M., & Mittone, L. (2010). Choice over sequences of outcomes: theory and experimental evidence. *Theory and Decision*, (69), 327–354.

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.

McCausland, W. J., Davis-Stober, C., Marley, A., Park, S., & Brown, N. (2020). Testing the random utility hypothesis directly. *The Economic Journal*, *130*(625), 183–207.

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

Nishimura, H. (2018). The transitive core: inference of welfare from nontransitive preference relation. *Theoretical Economics*, *13*, 579–606.

Ortoleva, P. (2013). The price of flexibility: towards a theory of thinking aversion. *Journal of Economic Theory*, (148), 903–934.

Reutskaja, E., Nagel, R., Camerer, C. F., & Rangel, A. (2011). Search dynamics in consumer choice under time pressure: an eye tracking study. *American Economic Review*, *101*, 900–926.

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.

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

Sen, A. (1997). Maximization and the act of choice. *Econometrica*, *65*(4), 745–779.

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.

Discussion Papers of the Research Area Markets and Choice 2023

Research Unit: **Market Behavior**

Levent Neyse, Frank M. Fossen, Magnus Johannesson, Anna Dreber SP II 2023-201
Cognitive reflection and 2D:4D: Evidence from a large population sample

Christian Basteck, Lars Ehlers SP II 2023-202
On the constrained efficiency of strategy-proof random assignment

Hande Erkut, Ernesto Reuben SP II 2023-203
Social networks and organizational helping behavior: Experimental evidence from the helping game

Research Unit: **Economics of Change**

Kai Barron, Tilman Fries SP II 2023-301
Narrative persuasion

Christina Timko, Maja Adena SP II 2023-302
Transparent app design reduces excessive usage time and increases willingness to pay compared to common behavioral design—a framed field experiment

Daniele Caliari SP II 2023-303
Behavioural welfare analysis and revealed preference: Theory and experimental evidence

All discussion papers are downloadable:
http://www.wzb.eu/en/publications/discussion-papers/markets-and-choice