Caliari, Daniele

On the relation between rationality and consistency

Social Choice and Welfare

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: Caliari, Daniele (2026) : On the relation between rationality and consistency, Social Choice and Welfare, ISSN 1432-217X, Springer Nature, Berlin, Iss. Online first articles, https://doi.org/10.1007/s00355-026-01656-8

This Version is available at:
https://hdl.handle.net/10419/339760

Social Choice and Welfare
https://doi.org/10.1007/s00355-026-01656-8

ORIGINAL PAPER

On the relation between rationality and consistency

Daniele Caliari[1]

Received: 15 January 2024 / Accepted: 19 February 2026

Abstract
We investigate whether the definition of economic rationality as choice consistency is correlated with decision-making ability. Guided by a theoretical framework, we demonstrate that documented positive correlations may be driven by confounding properties unrelated to consistency. To address this, in a novel experiment, we isolate consistency, measured by the Weak Axiom of Revealed Preference. We show that less sophisticated decision-makers often rely on simple rules and behave consistently, while more sophisticated ones consciously randomize, appearing inconsistent. These patterns determine ambiguous correlations, raising doubts about the choice of language that equates consistency with rationality in economics.

1 Introduction

> "Economists sometimes use the adjective rational in place of consistent, with the implied pejorative that choices that don't conform to their models are irrational. This is bad choice of language and is the source of all sorts of silly arguments with psychologists, sociologists, etc..." (Kreps 2015).

> "Rationality, they say, equals consistency... But this means only that those choices are consistent with one another *when viewed from the perspective of some theory* [italics in the original]" (Sugden 1991).

The theory under scrutiny is utility maximization: a decision-maker is a utility maximizer if and only if she has transitive and complete preferences, which is the case if

I am indebted to Marco Mariotti and Christopher Tyson for their advice and guidance. I also thank Georgios Gerasimou, David Freeman, Aniol Llorente-Saguer, Ivan Soraperra, Maria Vittoria Levati, David Dillenberger, Asen Ivanov, Tomas Jagelka, Dorothea Kubler, Lorenzo Neri, Pietro Ortoleva, Ariel Rubinstein, Steffen Huck, Yiming Liu, Kai Barron, Valentino Dardanoni, Itzhak Gilboa, and the participants of ESEM-EEA European Summer Meeting 2019, EEA Virtual Meeting 2020, WZB MC-Reading Group, and the first MPI-WZB workshop, 2021. I also thank Queen Mary University of London for funding the experiment and University of St. Andrews for hosting it. The experiment was approved by the Queen Mary Ethics of Research Committee: ref. QMREC2102. The open access publication was funded by the WZB Berlin Social Science Center.

✉ Daniele Caliari
daniele.caliari@wzb.eu

[1] Berlin Social Science Center, WZB, Berlin, Germany

and only if her choices are consistent. Economists have attached the adjective "rational" to consistent behaviour and documented correlations with proxies of decision-making ability such as wealth, education, and cognitive abilities (Burks et al. 2009; Choi et al. 2014; Andersson et al. 2016; Banks et al. 2019). The definition of consistency, however, is not innocuous. The classic notion (Arrow 1959; Sen 1971) is represented by the Weak Axiom of Revealed Preference (WARP), usually regarded as necessary for high-quality decision-making. Yet, several studies have employed stronger notions driven by a more structured theoretical framework (Choi et al. 2014) or by the experimental design (Burks et al. 2009; Andersson et al. 2016). In these cases, consistency has acquired a meaning beyond the mere existence of a utility function, encompassing properties that instead characterize its functional form. Some of these properties, e.g., monotonicity, may have driven the results, making it unclear whether the existing body of evidence justifies "the use of the adjective rational in place of consistent". In this paper, we take a step back and ask what these correlation exercises really capture.

Our opening argument becomes evident by comparing the prevalence of inconsistencies across some influential experiments in the literature. In Andersson et al. (2016), where Multiple Price Lists tie consistency to monotonicity, only 15% of the participants are inconsistent. The authors attribute inconsistencies to stochastic errors and, in line with this assumption, find a positive correlation between consistency and cognitive abilities. In contrast, in Agranov and Ortoleva (2017) and Rubinstein (2013), where monotonicity is not a confounding factor, the fraction of inconsistent participants spikes to 90% and 88%, respectively. Agranov and Ortoleva (2017) find that only 23% of the participants are inconsistent due to a stochastic error while 61% are so by a conscious choice of randomization. Rubinstein (2013), instead, finds that simple rules are an important driver of consistency questioning whether sophistication may be detrimental to consistent behaviour.

Neither Agranov and Ortoleva (2017) nor Rubinstein (2013) investigate the relationship between consistency and decision-making ability, which is our focus. We design an experiment that combines features of these studies while introducing some important novelties. More specifically, we induce the use of simple rules and randomization, measure cognitive abilities using a standard Raven test, and disentangle consistency from potentially confounding properties. Since this latter is a crucial distinction between our study and previous ones, we begin by zooming in on two well-known experimental designs in which consistency cannot be tested without further requirements on the utility function. These examples not only serve as motivation but will be analyzed from the lenses of a novel theoretical framework (Sect. 2) and a discussion of the existing literature (Sect. 7).

Example 1: In their influential paper titled "Who is (more) rational?", Choi et al. (2014) find that consistency measured by the Generalized Axiom of Revealed Preference (GARP) is correlated with wealth and education. By Afriat's Theorem (Afriat 1967), not only a utility maximizer with a strictly increasing utility function must satisfy GARP, but monotonicity is also necessary for GARP under locally non-satiated preferences. Consequently, in this setting, consistency has a stronger content than the mere existence of a utility function.

Example 2: Burks et al. (2009) and Andersson et al. (2016) use a classical Multiple Price List (MPL) design (Holt and Laury 2002; Andersen et al. 2008) to reveal a positive correlation between consistency and cognitive abilities. In this case, a failure of consistency is equated to multiple switches, and it implies a violation of first-order stochastic dominance: "We define subjects as Consistent if their decisions are compatible with rational [transitive and complete] and monotonic preferences" (Andersson et al. 2016). As in the previous example, consistency is inevitably tied to monotonicity, making it unclear how to test the former without the latter.

Our first step is to build a theoretical framework that identifies which properties represent consistency and which are, instead, confounding factors. We define two types of properties (henceforth axioms): consistency axioms (**ConAx**) and preference axioms (**PrAx**), which capture *how* the decision maker ought to choose, and *what* she ought to choose, respectively. We clarify our distinction by comparing WARP (**ConAx**) and monotonicity (**PrAx**) in the simplest possible setting. Consider a decision-maker who chooses between £5 and £6. On the one hand, WARP allows her to choose £5 over £6, but it restricts how subsequent choices are organized. Namely, **ConAx** are variations of the statement: "If you choose **a** when **b** is available in menu **A** then you should not choose **b** if **a** is available in menu **B**".[1] On the other hand, monotonicity imposes the choice of £6, in other words, **PrAx** are variations of the statement: "If **a** dominates **b** then you should not choose **b** when **a** is available." In our theoretical framework, we propose a symmetric feature of the axioms that disentangles **ConAx** and **PrAx**.

Guided by this taxonomy, in our design, the decision-maker faces two types of choice problems. In the MAIN problems, she can only violate **ConAx**; while in the remaining ones, violations of both **ConAx** and **PrAx** can arise. Similar to Burks et al. (2009) and Agranov and Ortoleva (2017), we focus on two choice domains: delayed payment plans (henceforth "Time") and lotteries (henceforth "Risk"), where the existence of a transitive and complete preference is a necessary condition for both discounted and expected utility theory.[2] Crucially, our MAIN problems are all the non-empty subsets of a set of four alternatives, making a test of WARP equivalent to a test of utility maximization (Sen 1971).

Figure 1 reveals an ambiguous and weak correlation between violations of WARP (in the MAIN problems) and Raven scores. This observation serves as the starting

[1] A review of different notions of consistency both deterministic and stochastic such as Independence from Irrelevant Alternatives, Stochastic Transitivity, and Regularity (Block and Marschak 1960), and their violations can be found in Rieskamp et al. (2006).

[2] If WARP is satisfied, then these theories are built on other **ConAx** such as stationarity and independence that also influence the shape of the maximized functions. However, these properties were not confounding factors in the existing literature, i.e., the positive correlations documented in Burks et al. (2009), Choi et al. (2014), and Andersson et al. (2016) are robust to relaxing both expected and discounted utility. In view of this, we defer their discussion to the Online Appendix.

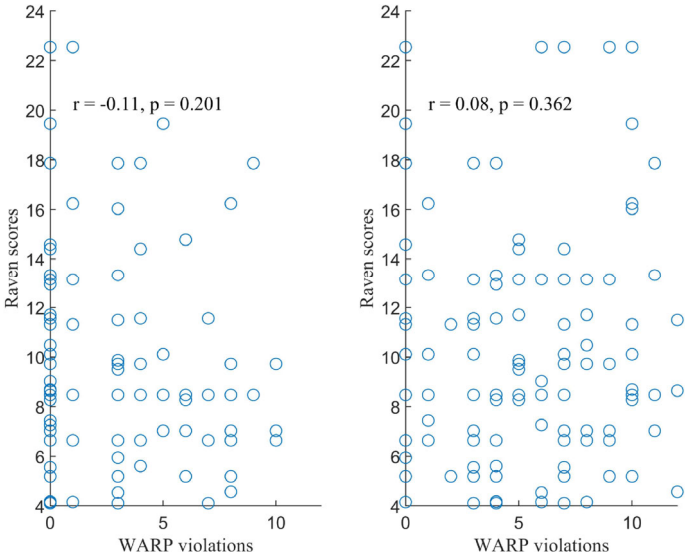

Fig. 1 Violations of WARP in Time (left) and Risk (right) and Raven scores

point for our analysis of the relation between consistency and cognitive abilities.[3]

We begin by assuming that the decision-maker maximizes a deterministic component, such as discounted (in Time) and expected (in Risk) utility, plus a stochastic error (Additive Random Utility Model [ARUM]). From this perspective, measuring consistency using WARP is natural because inconsistencies are interpreted as departures from deterministic utility maximization. Importantly, the conceptualization of noise as deriving from mistakes/errors is one of the two main interpretations of ARUMs (see Strzalecki 2025 for a discussion) and the one that is most common in the experimental literature relevant to our paper, e.g., Andersson et al. (2016).[4] Henceforth, we refer to this as the *noise-error* interpretation, which predicts a positive correlation between consistency and cognitive abilities—a pattern not observed in our experiment (Fig. 1). Our research question boils down to why.

We estimate individual preferences and precision parameters, reporting two facts: (i) a substantial fraction of participants have "extreme" preferences, and (ii) these participants are significantly more consistent than those with mild preferences. Within ARUM, one can be persuaded by the simple idea that extreme preferences, unlike mild ones, allow shocks in only one direction, leading to more consistent behavior.

[3] Despite its relevance, WARP is obviously not the only notion of consistency (see Sect. 7.3 for further discussions). It is possible that weaker notions of consistency can isolate inconsistencies equivalent to mistakes. We replicate our analysis using other indices of rationality, such as the Swaps index and the Houtman–Maks index, and consistency requirements from the literature (Manzini and Mariotti 2007), yet we still find weak correlations. Similarly, substituting Raven scores with Cognitive Reflection Test scores does not change this result. These robustness checks can be found in the Online Appendix.

[4] As Strzalecki (2025) notices, ARUMs have also been interpreted as unobserved preference heterogeneity, e.g. in Agranov and Ortoleva (2017). Though the two interpretations are equivalent in static environments, "they differ when it comes to normative evaluations" (Strzalecki 2025) as in our study.

We cannot exclude this mechanism, but we argue that noise alone cannot account for all observed patterns.[5] First, we note that the noise-error interpretation predicts a U-shaped relationship between preference parameters and Raven scores, a prediction that we reject. From here, a closer look at the participants' behaviour, i.e., their response times and violations of **PrAX**, reveals that for some of them the ARUM is misspecified. We show that the behaviour of participants with extreme preferences is better captured by heuristics, and in turn, that consistency is not a *sufficient* condition for high decision-making ability by validating Rubinstein's hypothesis (Rubinstein 2013): "Consistency may reflect the use of a simple rule rather than greater sophistication." While we find a strong positive correlation between extreme preferences and consistency in both Time and Risk, cognitive abilities only correlate with the complexity of the heuristics. We then investigate whether, among the remaining participants, the noise-error interpretation holds. We follow a growing literature (Agranov and Ortoleva 2017; Cerreia-Vioglio et al. 2019; Agranov and Ortoleva 2025) to show that inconsistencies often arise from "deliberate" choice. In doing so, we show that consistency is not even a *necessary* condition for high decision-making ability as participants who choose to randomize are significantly less consistent while displaying excellent results in cognitive abilities.[6] Finally, we show that for participants whose behaviour cannot be reconciled with heuristics or conscious randomization, the predicted positive correlation between consistency and cognitive abilities is confirmed, i.e., the noise-error interpretation applies.

2 Theoretical framework

A decision-maker faces a finite set of alternatives X and has preferences $\succeq$. $\mathcal{P}$ is the set of all possible preferences, $\mathcal{X}$ is the set of all non-empty subsets of X, $c : \mathcal{X} \to X$ is a choice function, and $\mathcal{C}$ is the set of all possible choice functions. Axioms (**Ax**) act by restricting the set of choices $\mathbf{Ax}(\mathcal{C}) \subseteq \mathcal{C}$. We characterize two types of axioms: consistency axioms (**ConAx**) and preference axioms (**PrAx**), using a symmetric property. Let $\pi : X \to X$ be a permutation over X. A permutation π acts on a choice function c such that for all menus A, $\pi c(A) = \pi(c(\pi^{-1}(A)))$. We say that an axiom is symmetric if for all π, $c \in \mathbf{Ax}(\mathcal{C})$ if and only if $\pi c \in \mathbf{Ax}(\mathcal{C})$; and it is asymmetric if this property is violated for some π.

ConAx are symmetric axioms while **PrAx** are not, a distinction that is reflected by the language in which these axioms are stated. For example, it is easy to see that WARP is a **ConAx** while monotonicity is a **PrAx**. WARP states that if $x = c(A)$ and $y \in A$ then $x, y \in B$ implies $y \neq c(B)$. Its definition contains no source of asymmetry; it only limits "how" choices are organized. Monotonicity, on the other hand, states that

[5] Overall, our study shows that consistency cannot be captured solely by a noise parameter that is a function of utility differences, but one (or more) other factor plays a role. This observation calls for a formal theoretical study, which is beyond the scope of this paper.

[6] These results are notably in contrast with Choi et al. (2014): "Some subjects may therefore adopt simple decision rules, and this "simplification" may cause their choices to be inconsistent."

if $x \succ_{\mathbf{PrAx}} y$ (for some order $\succ_{\mathbf{PrAx}}$) and $x, y \in A$ then $y \neq c(A)$. Clearly, $\succ_{\mathbf{PrAx}}$ provides a source of asymmetry that imposes "what" to choose irrespective of other choices.

The decision-maker behaves following a model, namely a mapping from $\mathcal{P}$ to $\mathcal{C}$. In the case of utility maximization the testable conditions are well-known (Sen 1971): a choice function satisfies WARP if and only if there exists a transitive and complete preference relation $\succeq$ that rationalizes it. The maximization model induces a mapping from axioms on $\mathcal{P}$ to axioms on $\mathcal{C}$ and vice versa,[7] i.e., WARP maps to transitivity and completeness while monotonicity on $\mathcal{C}$ maps to monotonicity on $\mathcal{P}$: if $x \succ_{\mathbf{PrAx}} y$ then $x \succ y$. This observation allows us to interpret the sole existence of a utility function as a **ConAx** and asymmetric properties of the utility function as confounding factors (**PrAx**).

We can now discuss our two motivating examples. The common theme is that in many relevant cases (multiple price lists, budget sets, etc...) consistency requirements imply a joint test of symmetric ("how to choose") and asymmetric ("what to choose") axioms.

Example 1 (continued): A decision maker faces the choice between n goods. $X = \mathbb{R}^n_+$ is the set of quantities, $p \in \mathbb{R}^n_{++}$ is a price vector, I is the income, and $x \in X$ is a consumption bundle. A budget set $B(p, I)$ is the set of bundles such that $p \cdot x \leq I$. If a bundle x_i is chosen at prices p_i, we write (x_i, p_i) and we say that $x_i \succeq (\succ) x_j$ if $p_i \cdot x_j \leq (<) p_i \cdot x_i$. GARP can be simply stated as follows: if $x_i \succeq x_j$ then there exists no sequence $x_1 \ldots x_n$ such that $x_j \succeq x_1 \ldots x_n \succeq x_i$ with one strict preference. GARP seems nothing more than a **ConAx**.[8] However, in this setting, not only does monotonicity imply GARP, but under local non-satiation, Afriat's Theorem implies that the converse also holds: observed violations of monotonicity, which is testable in the data, unlike concavity, imply violations of GARP. This observation implies that **ConAx**, interpreted as the existence of a utility function, cannot be tested without also testing **PrAx**.

Example 2 (continued): A multiple price list is a sequence of binary sets $(\{x_i, y\})_{i=1}^n$ that compares the same alternative y to a series of alternatives x_i such that for all $i < j$, $x_i \succ_{\mathbf{PrAx}} x_j$ for some order $\succ_{\mathbf{PrAx}}$. Namely, as we progress in the multiple price list the first alternative becomes worse according to $\succ_{\mathbf{PrAx}}$. In Andersson et al. (2016), X is a set of lotteries and $\succ_{\mathbf{PrAx}}$ is the order representing first-order stochastic dominance (FOSD). Imagine observing a multiple switch such that $y = c(x_i, y)$ and $x_j = c(x_j, y)$ with $j > i$. If transitivity is assumed then FOSD is violated: $x_j \succ y \succ x_i$ implies $x_j \succ x_i$ implying again that **ConAx** cannot be tested separately from **PrAx**.

[7] See Mahmoud (2017) for a study of the mapping between axioms on primitives and observables.

[8] Notice that some **PrAx** are built into the structure of the problem. The definition of the revealed preference $\succeq$ immediately implies Walras' Law, which in turn is a result of the problem being defined among goods (not bads) and local non-satiation. Note that this latter is crucial for testing the hypothesis of utility maximization, as otherwise unstructured utility functions, such as a constant utility, can trivially represent the data.

Table 1 List of the 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

Notes: The amounts are described in Token with an exchange rate of 20:1 pounds

Table 2 List of the MAIN lotteries

Alternatives	Token		Probabilities		EV
Degenerate (D)	50	0	1.0	0.0	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 with an exchange rate of 10:1 pounds

3 Our theoretical environment and the experiment

Our experiment has a within-subject design. In Parts One and Two, participants were asked to choose from different sets of alternatives: delayed payment plans (Time) and lotteries (Risk). A delayed payment plan, $d \in D$, is a tuple $(m_1, \ldots, m_5; t_1 \ldots t_5)$ with $m_1, \ldots, m_5$ being monetary prizes and $(t_1, \ldots, t_5) = (0, 3, 6, 9, 12)$ being the months after which the respective payments are received. A lottery, $l \in L$, is a tuple $(x_1, x_2; p_1, p_2)$ with $x_1, x_2 \in \mathbb{R}$ being monetary prizes and $p_1, p_2 \in [0, 1]$ being the probabilities respectively of x_1, x_2.

Each part had 25 choice problems designed to be non-trivial. In both Time and Risk, we focus on a subset of four alternatives (see Tables 1 and 2) and collect choices on all the eleven non-empty subsets of these alternatives with cardinality at least two. We refer to these sets as the MAIN problems and to the four alternatives as the MAIN alternatives (see the Online Appendix for a complete description of the alternatives).

We measure (in)consistency focusing on the number of violations of WARP (**ConAx**), namely the sum over all pairs of elements of the product between the number of times an element x is chosen when y is available, and vice versa. Formally, for all $x, y \in X$, let C_{xy} be the number of times x is chosen when y is available (see the Online Appendix for robustness checks to our measurement of consistency):

$$\text{WARP violations} = \sum_{x,y} C_{xy} \cdot C_{yx}$$

Guided by the literature, we focus on two **PrAx** in Time and two in Risk that have characterized some influential experiments in the past, and have been potential confounding factors for the study of consistency. The first two **PrAx** are the most basic

and fundamental, and both in Time and Risk describe a monotonic preference over the monetary outcomes: (i) Monotonicity (MON): for all $d, \hat{d}$, we say $d >_{\mathbf{MON}} \hat{d}$ if for all $i = 1, \ldots, 5$, $m_i \geq \hat{m}_i$ with at least one strict inequality; (ii) First Order Stochastic Dominance (FOSD): for all $l, \hat{l}$, we say $l >_{\mathbf{FOSD}} \hat{l}$ if given F_l and $F_{\hat{l}}$ the respective cumulative distribution functions: $F_{\hat{l}}(x) \geq F_l(x)$ for all $x \in \mathbb{R}$. These monotonic axioms are typical confounders in Multiple Price Lists designs such as, among many, Andersen et al. (2008), Burks et al. (2009), and Andersson et al. (2016). The second two **PrAx**, instead, are constraints on the preferences over the time and risk dimensions: (i) Impatience (IMP): for all $d, \hat{d}$, we say $d >_{\mathbf{IMP}} \hat{d}$ if the vector $\hat{m}$ is a permutation of m and the non-zero differences $m_i - \hat{m}_i$ are first positive and then negative, with at most one sign change[9]; (ii) Second Order Stochastic Dominance (SOSD): for all $l, \hat{l}$, we say $l >_{\mathbf{SOSD}} \hat{l}$ if $\int_{-\infty}^{x} [F_{\hat{l}}(h) - F_l(h)]dh \geq 0$ for all $x \in \mathbb{R}$. These last axioms have also played an important role in the experimental literature. For example, in the influential experiment of Tversky and Russo (1969), repeated by McCausland et al. (2019), alternatives were ranked by SOSD; while in Manzini et al. (2010) alternatives were ranked by IMP.

Within the MAIN problems, participants could not violate **PrAx**, as it is clear from Tables 1 and 2. Therefore, from now on, when referring to (in)consistency, we mean WARP violations within the MAIN problems where confounding properties play no role. Beyond the MAIN problems instead both **PrAx** and **ConAx** could be violated, and alternatives were constructed such that each MAIN alternative is matched to a dominated alternative for each **PrAx** (see the Online Appendix for details).

At the beginning of the experiment, participants received general instructions plus specific instructions about both parts.[10] Furthermore, before Parts One and Two, participants answered three trial problems in order to make them familiar with the experimental design. The positions of the alternatives were randomized. The participants could face two orders of problems, and we also inverted Time and Risk elicitation for a total of four treatments. A complete structure of the experiment and a description of the orders can be found in section 2 of the Online Appendix. At the end of the experiment, participants answered (non-incentivized) a subset of ten Raven matrices. Finally, one choice was randomly selected in both Time and Risk and paid out. The reward was measured in tokens with an exchange rate of 1:10 for lotteries and 1:20 for delayed payment plans. The average reward was approximately £19, and the experiment lasted on average 1 h and 15 min. The experiment took place at the University of St. Andrews between June and September 2019, with a sample of 145 participants, and it was conducted using z-tree (Fischbacher 2007).

[9] We borrow the term "impatience" from Fishburn and Rubinstein (1982) where it is denoted as Axiom A3.

[10] The instructions were available on screen and paper so that they could be consulted during the entire experiment. Note that the experiment has been designed to be paper-free.

4 Interpreting (in)consistency

4.1 The noise-error interpretation

We first consider a decision maker who maximizes a utility function, but she does so with a stochastic error (ARUM). Let U_θ be a family of utility functions with parameter θ. Within our main analysis, in Time, U represents the exponential discounting model with $\theta = \beta$ as the discount factor, while in Risk, it represents the CRRA utility function with $\theta = \rho$ as the risk aversion parameter.[11] In ARUMs, the probability of choosing an alternative x from a set A is $p_\theta(x, A) = Prob\{U_\theta(x) + \varepsilon(x) \geq U_\theta(y) + \varepsilon(y)$ for all $y \in A\}$. More concretely, stochasticity is modelled using a logit function with the precision parameter λ such that $p_\theta(x, A) = \frac{e^{\lambda U_\theta(x)}}{\sum_{y \in A} e^{\lambda U_\theta(y)}}$. Due to our interest in heterogeneous behaviour[12] we estimate for each participant i, the discount factor β_i, the risk aversion parameter ρ_i, and two precision parameters λ_i in Time and Risk. In Figs. 2 and 3, we summarize our estimates by plotting the joint distributions of preference and precision parameters. The main observations are: (i) a substantial fraction of participants have "extreme" preferences, and (ii) these participants are significantly more consistent than those with mild preferences.

An interpretation of noise involving errors implies that cognitive abilities should be positively correlated with the estimated individual precision (a proxy of consistency, i.e. the Spearman correlation between precision and WARP violations is -0.78 in Time and -0.63 in Risk), and have a U-shaped relation with the estimated preference parameters. We reject both predictions. First, we replicate the results from Fig. 1 as the (Spearman) correlation between Raven scores and precision is 0.06 in Time (p-value = 0.464) and -0.04 in Risk (p-value = 0.613).[13] Second, both in Time and in Risk we find a monotone (not U-shaped) relation between preference parameters and cognitive abilities ($+0.15$ in Time, p-value = 0.059, and -0.15 in Risk, p-value = 0.062). In the following sections, we will interpret this evidence from different lenses.

4.2 Extreme preferences: failures of the ARUM

Extreme preferences are clearly a key driver of consistency and deserve a detailed analysis. We categorize these participants using the top and bottom deciles of the parameter space.[14] We call **impatient** those with discount factor: $0.9 < \beta < 0.91$; **patient**: $0.99 < \beta < 1$; **risk-averse**: $2.2 < \rho < 2.4$; and **risk-neutral**: $0 < \rho < 0.2$.

[11] Since our analysis relies on the correct specification of cardinal utilities, we provide several robustness checks. More specifically, in Appendix 1, we estimate the model using different specifications, e.g., CARA, and certainty equivalents in Risk, and Hyperbolic and Quasi-hyperbolic discounting in Time. In the online appendix, we present alternative model specifications that do not rely on cardinal assumptions and an alternative specification of the error, the Random Parameter Model (Apesteguia and Ballester 2018). We confirm the results.

[12] The full details regarding the structural model and its estimation (Train 2008) are in Appendix C.

[13] Given that the non-linearity of the logit function is reflected in the precision parameter, we report the Spearman correlation coefficient. Results do not change using linear correlations.

[14] In the Online Appendix, we show that results are robust when categorizing participants using different quintiles.

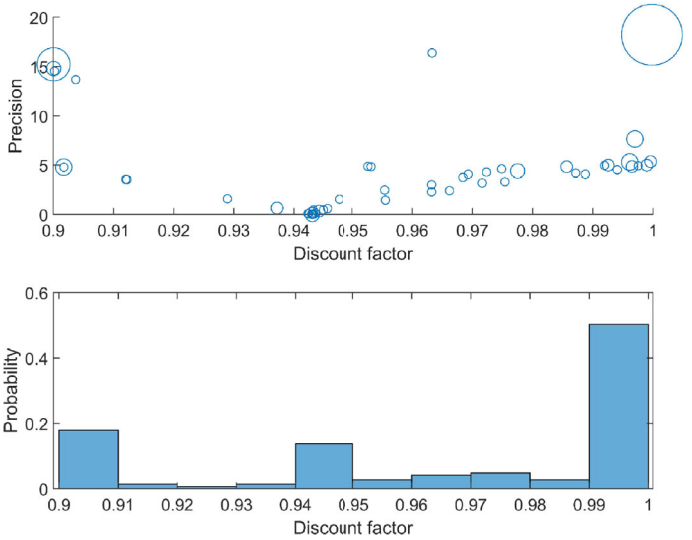

Fig. 2 On the top, the joint distribution of β and λ. On the bottom, the marginal distribution of β. Notes: the top panels report the joint distribution of individual preference and precision parameters. The size of the bubble is proportional to the number of participants. Precision takes different values in Time and Risk due to the different scales of the problems so comparisons are only meaningful within domains. The bottom panels report the marginal distributions on the preference parameter space

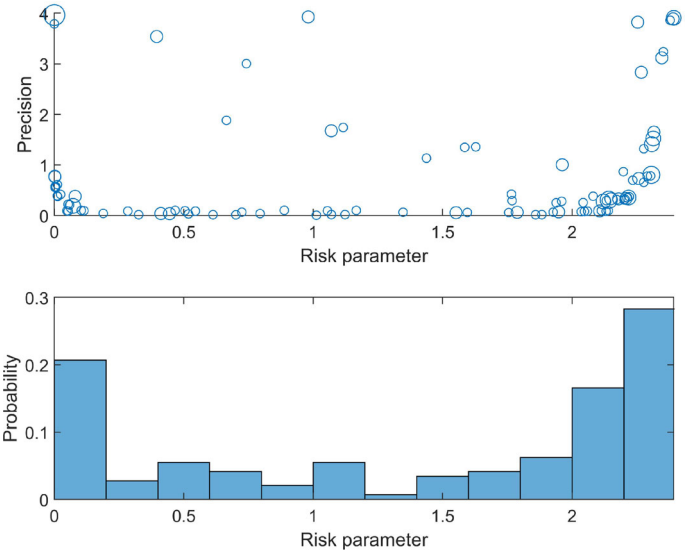

Fig. 3 On the top, the joint distribution of ρ and λ. On the bottom, the marginal distribution of ρ. Notes: the top panels report the joint distribution of individual preference and precision parameters. The size of the bubble is proportional to the number of participants. Precision takes different values in Time and Risk due to the different scales of the problems so comparisons are only meaningful within domains. The bottom panels report the marginal distributions on the preference parameter space

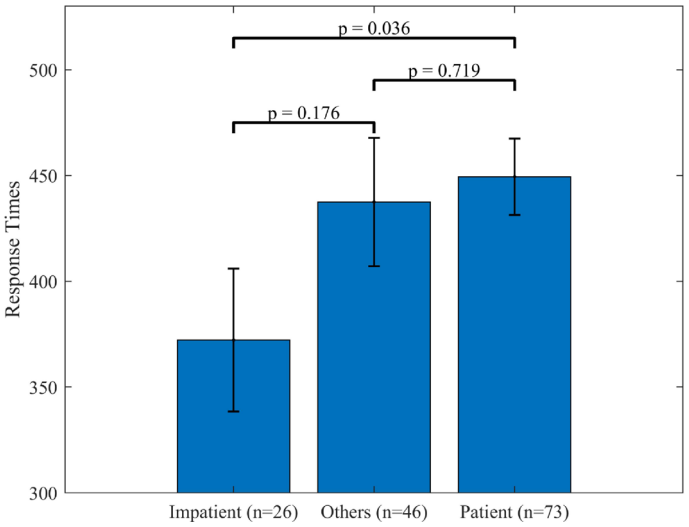

Fig. 4 Response times in Time. Notes: The bar plots report the sum of the response times in all questions and the participants are grouped by their preference parameter. On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests

Comparing the behaviour of participants with mild and extreme preferences reveals two failures of the ARUM. First, we test the common assumption known as the chronometric function, i.e., response times are decreasing in utility differences, or equivalently for us, participants with extreme preferences should be quicker (see Alós-Ferrer et al. 2021).[15] Figures 4 and 5 reject this assumption, showing a clear monotonic relation between preferences and response times, with impatient and risk-averse participants being significantly quicker than patient and risk-neutral ones.

The second failure of the ARUM is observed in the probability of violating IMP and SOSD, which should naturally be monotone in the preference parameters. Figures 6 and 7, instead, show a clear non-monotonicity. In other words, patient participants are less likely to violate IMP than mildly impatient ones, while risk-neutral participants are less likely to violate SOSD than mildly risk-averse ones. Importantly, this result is not driven by the higher consistency of participants with extreme preferences, as it is confirmed if we restrict our focus to participants with mild preferences who are, on average, as consistent as those with extreme ones.

The stark evidence of Figs. 4, 5, 6 and 7 has a heuristic-based interpretation. The framing of our experiment induced heuristics that map into extreme preferences. Lot-

[15] The prediction that participants with extreme preferences should be quicker follows three observations. First, the chronometric function implies that response times are decreasing in utility differences. Second, since utility differences are non-monotone along the parameter space (see Wilcox 2011; Apesteguia and Ballester 2018 for a discussion), we assume that the chronometric assumption holds on the normalized utility space (Wilcox 2011). Third, once utilities are normalized, the sum of the utility differences is larger for participants with extreme preferences. Finally, note that Figs. 2 and 3 suggest that utility differences for these participants are even larger than those captured by our experimental design, as most are on the boundary of the parameter space. This observation reinforces the idea that their response times should be lower.

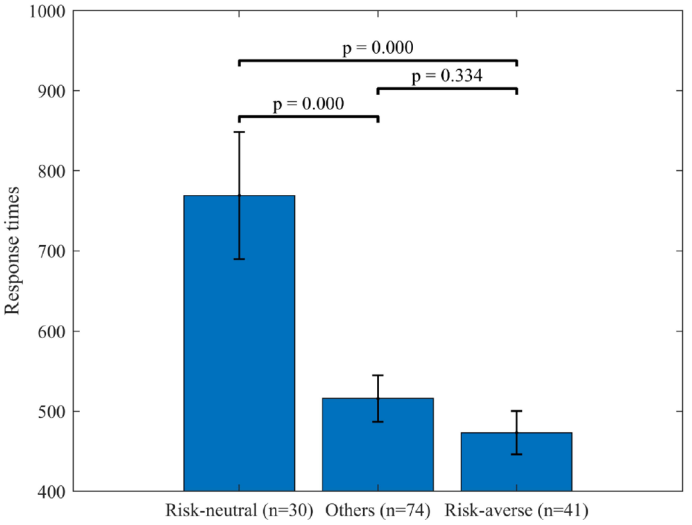

Fig. 5 Response times in Risk. Notes: The bar plots report the sum of the response times in all questions and the participants are grouped by their preference parameter. On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests

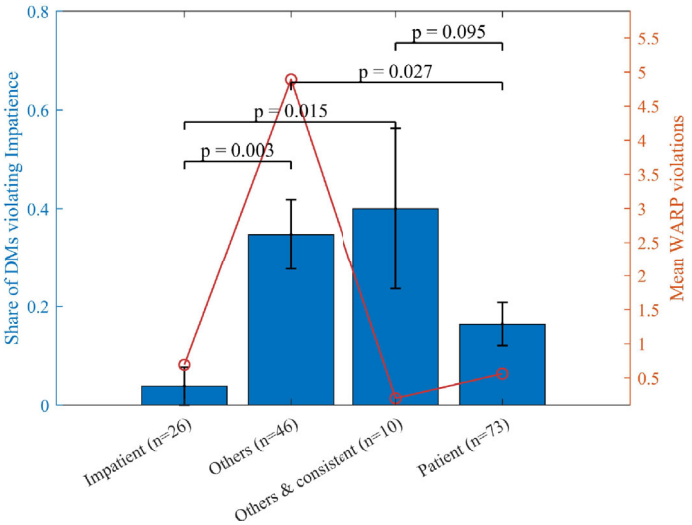

Fig. 6 Share of participants violating IMP in Time. Notes: The bar plots report the share of participants who violate IMP within each group defined by preference parameters. The group "consistent" is defined as those participants with WARP violations lower or equal than x, where x is found such that the overall mean in the group is the closest to that of the participants with extreme preferences. The plot reports the average number of WARP violations in each group. On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests

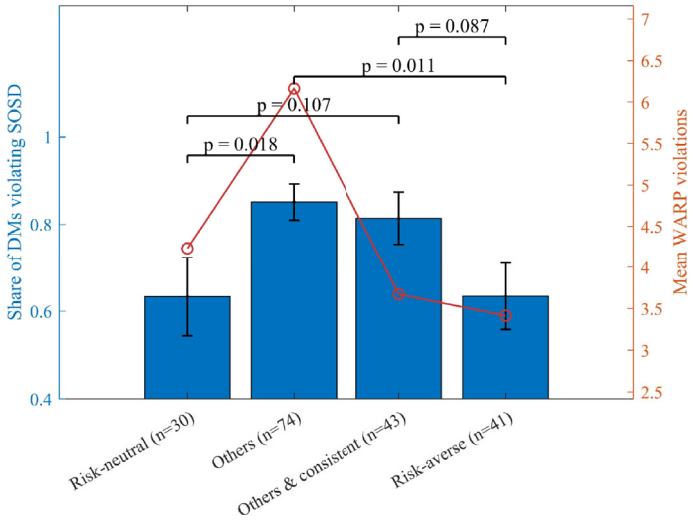

Fig. 7 Share of participants violating SOSD in Risk. Notes: The bar plots report the share of participants who violate SOSD within each group defined by preference parameters. The group "consistent" is defined as those participants with WARP violations lower or equal than x, where x is found such that the overall mean in the group is the closest to that of the participants with extreme preferences. The plot reports the average number of WARP violations in each group. On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests

teries were represented by pies and delayed payment plans by histograms, and both were complemented by tables reporting probabilities/timings and monetary amounts. This framing implies that, on the one hand, **risk-averse** and **impatient** participants could apply simple visual heuristics, such as searching for the highest probability of the highest amount (Risk) and the highest first histogram (Time). On the other hand, **risk-neutral** and **patient** participants followed heuristics that involved calculations such as choosing the highest expected value (Risk) and the sum of all monetary amounts (Time). To rationalize our evidence, notice that the different complexities of the heuristics induced the differences in response times observed in Figs. 4 and 5,[16] while the consistency induced by the use of heuristics explains the non-monotonic behaviour observed in Figs. 6 and 7.[17]

[16] This evidence, under a similar framing, appears also in Agranov et al. (2025): "safe choices seem to be related to intuitive, less thoughtful choice processes...risky option seems to require a more deliberate decision".

[17] Note also that **impatient** participants by choosing the first and highest histogram have a perfect remedy against violations of IMP and, in fact, only 1 out of 26 of them violate this axiom (Fisher exact test with the group of all other participants, violations 28 out of 119, p-value = 0.028). **Patient** participants, instead, by summing all the payments, have the perfect remedy against violations of MON, and again, only 2 out of 73 of them violate this axiom, a proportion that is significantly lower than any other group (Fisher exact test with the group of all other participants, violations 11 out of 61, p-value = 0.009).

4.3 Are inconsistencies errors?

So far, we have shown that participants with extreme preferences likely use heuristics to simplify the choice process. We now ask whether we can interpret errors as simply noise or whether other factors may play a role. Following Agranov and Ortoleva (2017), we use the questionnaire (in the Online Appendix we list a sample of the participants' answers) to show that a substantial fraction of participants consciously randomize (henceforth **Randomizers**) rejecting errors as sole driver of the observed noise.[18,19]

Beyond the reported behaviour, we provide two additional pieces of evidence in favour of our identification of randomizers. The first is based on replicating the findings of Agranov and Ortoleva (2017), showing that randomization is significantly more widespread in Risk than in Time (19% vs 5%, McNemar test, p-value = 0.000). This observation further explains two facts: (i) 69% of participants have extreme preferences in Time while only 37% in Risk as randomizers, in a potentially misspecified ARUM, display mild levels of risk-aversion; (ii) the average number of WARP violations is significantly higher in Risk than in Time (4.99 vs 1.96, unpaired t-test, p-value = 0.000). The second piece of evidence exploits the possibility, in our design, of violating both FOSD and SOSD outside the MAIN problems. As described in Appendix A, Cerreia-Vioglio et al. (2019) show that participants who randomize should not choose first-order stochastically dominated alternatives, while there is no constraint on second-order stochastically dominated ones. This implies that (i) randomizers should violate SOSD more often, and (ii) we should observe a positive correlation between violations of SOSD and inconsistencies within the MAIN problems. In line with these predictions, we find (i) 90% of randomizers violating SOSD, while this proportion is 71% among the remaining participants (Fisher exact test, p-value = 0.054); and (ii) a strong positive correlation between violations of SOSD and WARP violations (r = 0.30, p-value = 0.000).[20]

5 Main result: rationality and consistency

Recall, within the noise-error interpretation, if individuals with higher cognitive abilities make fewer mistakes, consistency is logically linked to cognitive ability. In previous sections, we have shown that the noise-error interpretation alone cannot explain the choices of a subset of participants, implying that this logical link may be

[18] Examples of randomization in choices among lotteries are Hey and Carbone (1995), Ballinger and Wilcox (1997), Sopher and Narramore (2000), Hey (2001), Agranov and Ortoleva (2017), Yu et al. (2021) Evidence of randomization between delayed payment plans can be found in studies involving Multiple Price Lists such as Andersen et al. (2008).

[19] In the online appendix, we provide another evidence that failures of ARUM documented in the previous section are driven by participants who are likely to either use heuristics or to randomize. Specifically, we exclude these participants and rerun our analysis in Sect. 4.2, finding results closer to the predictions of ARUM.

[20] The correlation between violations of WARP and SOSD does not depend on the use of heuristics. When focusing on participants with non-extreme preferences, the correlation is still significant (r = 0.27, p-value = 0.049).

absent. Here, we show that this is indeed the case. First, our interpretation through the lens of the complexity of the heuristic translates into a reformulation of Rubinstein's hypothesis (Rubinstein 2013): "Consistency with transitivity may reflect the use of a simple rule rather than greater sophistication".

Hypothesis 1 Participants who use heuristics are (i) more consistent but (ii) not more sophisticated than other participants, especially when adopting simple rules.

Second, (conscious) randomizers, as those modelled by Cerreia-Vioglio et al. (2019), cannot be narrow bracketers, i.e. their future choices depend on the old ones. This observation implies that decision-makers with high cognitive abilities may choose inconsistently and leads to the following hypothesis.

Hypothesis 2 Randomizers are (i) less consistent but (ii) not less sophisticated than other participants.

Figure 8 reveals that participants who use heuristics (simple or not) are significantly more consistent than the remaining ones, confirming (i) in Hypothesis 1, while Fig. 9 shows no evidence of a relationship with Raven scores.[21] In detail, we do not reject the null hypothesis that participants who use heuristics are not more sophisticated, confirming (ii) in Hypothesis 1.

Figures 8 and 9 also confirm Hypothesis 2. In Risk, randomizers are more likely to be inconsistent,[22] but they are no less sophisticated. In fact, we even reject the alternative hypothesis that randomizers are less sophisticated (p-value = 0.077 w.r.t. other participants, p-value = 0.043 w.r.t. participants who rely on simple heuristics that yield risk-averse preferences).[23]

5.1 The relation between rationality and consistency: heterogeneous analysis

To stress the importance of choosing the correct model of noise when interpreting the correlation between consistency and cognitive abilities, we conclude with some thought-provoking figures. Using the same groups defined in the previous section, in Fig. 10, we plot the correlation between consistency and cognitive abilities, partitioning our sample into two groups: (i) extreme and randomization; and (ii) others.

[21] The relation between preferences and cognitive abilities has been the subject of extensive research, see e.g. Andersson et al. (2016) for a review on risk preferences. In this paper, we cannot disentangle whether there is an inherent relationship between preferences and cognitive abilities or whether complexity can explain past results. There is currently a lively literature on complexity, see e.g. Oprea (2024), Agranov et al. (2025), and we believe this question is worth investigation.

[22] There is a small overlap between participants who are coded as randomizers and have extreme preferences. Not surprisingly, we find that these participants (n = 10) are significantly more likely to violate WARP than those who have extreme preferences but are not coded as randomizers (p-value = 0.002).

[23] We provide several robustness checks to the analysis in this section. In Appendix 1, we use reported behaviours from the questionnaire to identify the use of heuristics. We find strong correlations with extreme preferences and replicate the patterns of Figs. 8 and 9. In Appendix B.2, we reveal the participants' preferences using non-parametric methods such as the Minimum Swaps algorithm (Apesteguia and Ballester 2015), and the Sequential algorithm (Horan and Sprumont 2016), and again replicate our findings. Finally, in the Online Appendix, we replicate our analysis using several other approaches.

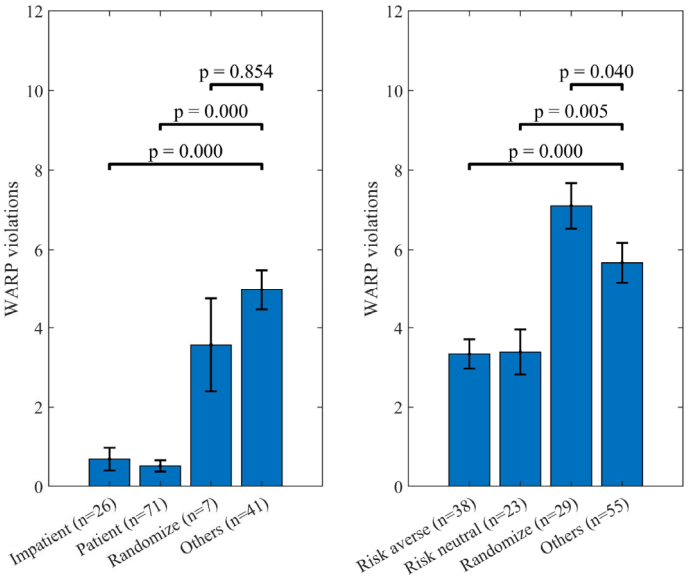

Fig. 8 WARP violations (heterogeneous analysis). Notes: The two plots report the average number of WARP violations for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests. Importantly, since our hypotheses are one-directional, the statistical tests are one-tailed

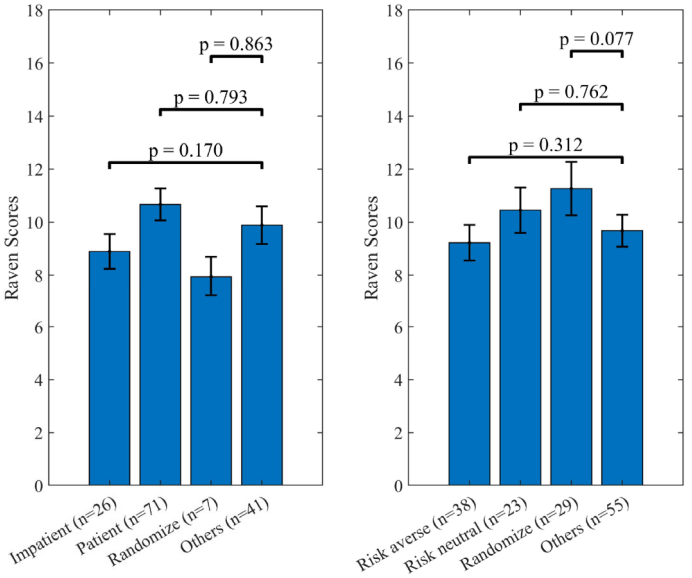

Fig. 9 Raven scores (heterogeneous analysis). Notes: The two plots report the average Raven scores for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests. Importantly, since our hypotheses are one-directional, the statistical tests are one-tailed

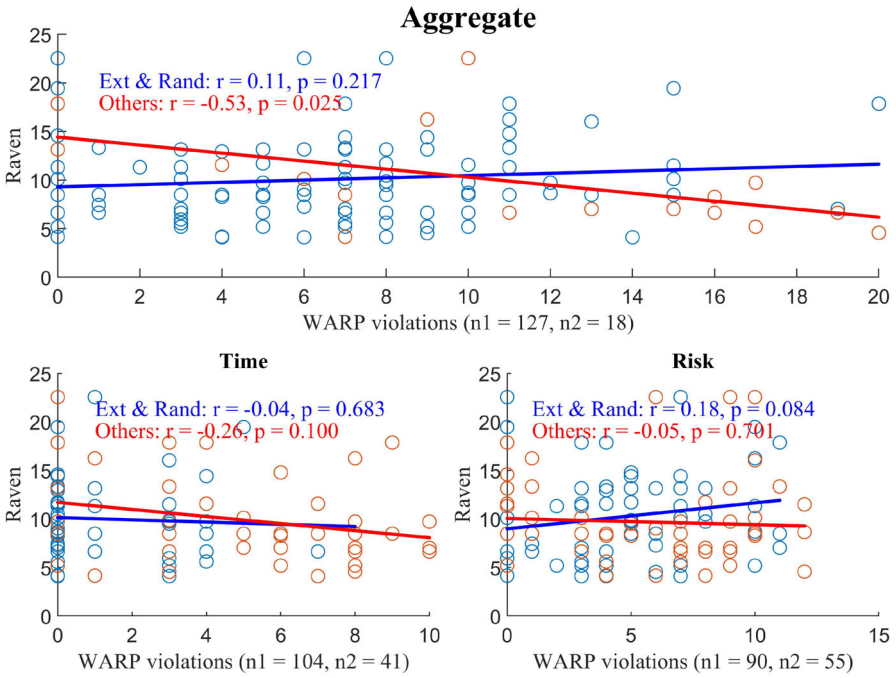

Fig. 10 Correlation between WARP violations and Raven scores (heterogeneous analysis). Notes: the three panels report the joint distributions of Raven scores and WARP violations for two categories of participants. "Ext & Rand" are either participants with extreme preferences or randomizers. "Others" are the remaining participants. In the top panel, we denote "Ext & Rand" participants who are categorized as such either in Time or in Risk. In the parentheses, we report the numerosity of the groups with $n1$ representing "Ext & Rand" and $n2$ representing "Others"

The main panel shows data aggregated by Time and Risk, where the "others" group is defined as those participants who are never coded as having extreme preferences or as randomizers. The smaller panels focus separately on Time and Risk. As expected, we find different correlation patterns in the two groups, mostly with opposite signs (Fisher z-tests: in aggregate p-value = 0.011, in Time p-value = 0.235, in Risk p-value = 0.186). These results speculatively suggest that the noise-error interpretation of ARUMs for the "others" group is adequate.

6 Further results: preference axioms and consistency

Before our final discussion, we investigate whether **PrAx**, in our experiment, play the role of confounding factors. First, we look at the prevalence of **PrAx** violations and compare these results with the literature. In Time, 9% of participants violated MON while 20% violated IMP, and these violations are correlated (r = 0.27, p-value = 0.001). In Risk, 26% of the participants violated FOSD while 74% violated SOSD, with a lower correlation between them (r = 0.1, p-value = 0.246), confirming that the sources of violations of these two axioms are different. Our results are similar to

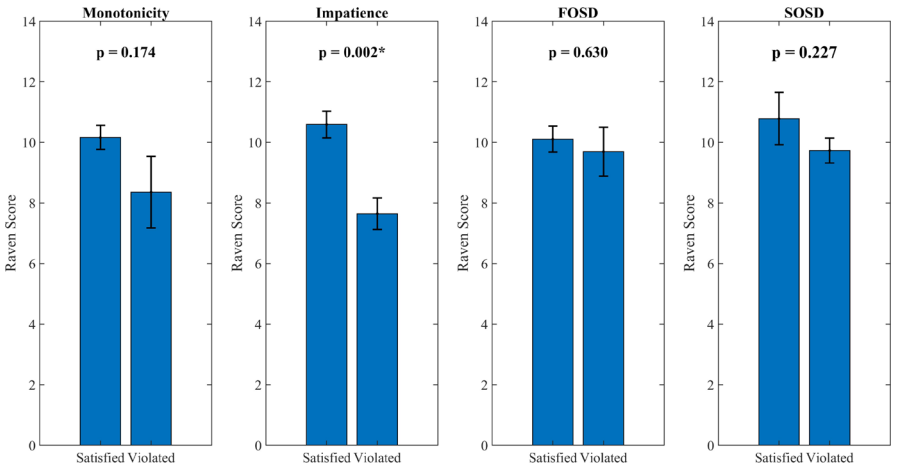

Fig. 11 Violations of **PrAx** and Raven scores. Notes: The bars show the average Raven scores for participants who either satisfy or violate different **PrAx**. Above the bars, we include the p-values from Welch's t-tests

Burks et al. (2009) with significantly more violations of monotonicity in Risk than in Time (McNemar test, p-value = 0.000) and furthermore, in Risk, the prevalence of violations of FOSD resemble the proportion of multiple switching in the literature (Filippin and Crosetto (2016), see Sect. 7 for a further discussion).

Figure 11 shows the correlation between violations of **PrAx** and cognitive abilities. We find negative correlations everywhere, though only significant for IMP. Noticeably, since patient participants rarely violate MON due to their heuristic (see footnote 17), among the remaining participants, the correlation is larger and significant (p-value = 0.026). The results in Risk are less clear. In particular, while violations of SOSD can be rationalized by appealing to randomization, we fail to explain the absence of correlation regarding violations of FOSD and we do not find any significant effect of extreme preferences or randomization for this property.

Finally, we ask whether participants who violate **PrAx** overlap with those who violate **ConAx**. The stark differences in the prevalence of violations, i.e., excluding SOSD, 42% of participants violate one of MON, IMP, and FOSD while 88% violate WARP at least once, suggest that a substantial group of participants violate WARP but not monotonicity (in any form). Table 3 shows only weak correlations that become even weaker once controlling for the use of heuristics, as they induce higher consistency both for **ConAx** and **PrAx**. The only exception is SOSD, for which we find a strong and positive correlation that drops after controlling for randomization, in line with our discussion in the previous section.

7 Discussion and further research

Our study is naturally related to the literature on GARP and multiple switching behaviour (MSB). Here, we build on our motivating examples to show how our find-

Table 3 Correlation between violations of **ConAx** and **PrAx** in Time

	WARP Time			WARP Risk		
	(1)	(2)	(3)	(1)	(2)	(3)
	Panel A—Impatience			Panel A—FOSD		
	0.784	−0.499	−0.302	0.411	0.489	0.412
	(0.587)	(0.518)	(0.512)	(0.647)	(0.588)	(0.574)
	Panel B—Monotonicity			Panel B—SOSD		
	0.806	−0.277	−0.160	2.376***	1.537**	1.228*
	(0.796)	(0.488)	(0.499)	(0.610)	(0.600)	(0.652)
Extreme and random	×	✓	✓	×	✓	✓
Response times	×	×	✓	×	×	✓
Observations:	145	145	145	145	145	145

Notes: The dependent variable is the number of violations of WARP, and the main independent variables are the dummy variables representing violations of **PrAx**. Column 1 displays the unconditional correlations. In column 2, we control for dummy variables representing extreme preferences and randomization. Finally, in column 3, we control for response times within ALL problems. The regression models are estimated in Stata (Robust standard errors). ***< 0.01, **< 0.05, *< 0.1

ings fit into the existing literature, and we provide further theoretical mechanisms that could explain the differences. Then, we discuss the assumptions behind our analysis and, in particular, the focus on WARP and utility maximization.

7.1 Discussion: violations of GARP

As mentioned in the introduction, **ConAx** are often considered as necessary conditions for high-quality behaviour as expressed by Choi et al. (2014): "if decisions are high-quality then there exists a utility function the choices maximize" (see also Kariv and Silverman 2013; Carvalho and Silverman 2019). The authors (see Section II.B in Choi et al. 2014) highlight that GARP does not impose normative restrictions on the utility function showing that GARP allows for violations of FOSD. We believe this point is worth a discussion in view of our findings.

Choi et al. (2007, 2014), and Dembo et al. (2022) find that the vast majority of participants violate GARP, even if mildly, a result that may be expected given the high number of choices participants encountered in their experiment (see Figure 4, in Choi et al. 2007). However, in these papers, not only do most of the participants violate GARP, but they also violate FOSD, i.e. Dembo et al. (2022) find that no participant satisfies FOSD in a 3-dimensional case, while only a handful satisfy it in the 2-dimensional case (see the Appendix in Dembo et al. 2022, and similarly Choi et al. 2007, 2014). The high number of both GARP and FOSD violations seems to be a strikingly different result from that observed in other studies. For instance, in Agranov and Ortoleva (2017) and in our experiment, almost all participants violate WARP (90% and 85.5%, respectively) but much less violate FOSD (6% and 26%, respectively). To reconcile this evidence, we propose the following explanation. First, in the context of budget sets, FOSD is a complex theoretical concept that involves interactions

among states of the world (see Dembo et al. 2022). GARP, instead, imposes constraints (monotonicity and concavity) only within states. Given this premise, the high number of GARP violations is consistent with the high number of WARP violations observed in choice under risk in Agranov and Ortoleva (2017) and in our experiment. On the other hand, the low prevalence of FOSD violations in these latter studies can be rationalized by the fact that states of the world are unlabeled. Here, FOSD implies the existence of a probability space and two random variables ordered by state-by-state stochastic dominance (see Theorem 1 in Hansen et al. 1978 and subsequent discussions), i.e. participants may be able to apply FOSD when it overlaps with more intuitive notions of stochastic dominance, while they fail when this is not the case, as in Choi et al. (2014) and Dembo et al. (2022). We believe further research is needed to deepen our understanding of why decision-makers may violate (or satisfy) different stochastic dominance properties, as well as other **PrAx**, and how these properties are related to **ConAx**.

7.2 Discussion: multiple switching behaviour

A mirroring discussion can be developed focusing on the literature on MSB. As anticipated, MPLs do not allow for a separate test of **ConAx** and **PrAx**. This is strikingly clear in the existing literature. Filippin and Crosetto (2016) collect 54 studies involving 6315 participants and find an average of 17.1% of participants display MSB. In line with this prediction, Andersson et al. (2016) find 14.8% and 30.48% in their two treatments, respectively. These numbers are interestingly aligned with the proportion of our participants who violate FOSD, but in stark contrast to the proportion who violate WARP.

More recently, Chew et al. (2022) introduced a distinction between regular and irregular MSB (see Chew et al. 2022 for details), arguing that only the former can be explained by conscious randomization. The authors then look at the correlation between cognitive abilities and MSB and find that both participants with regular and irregular MSB display a negative correlation, but with the latter having a regression coefficient twice that of the former (see Table 3 in Chew et al. 2022). Since the authors use randomization to explain their findings, this evidence appears to support our results. However, the result does not seem robust. First, using data from Andersson et al. (2016), we confirm a negative correlation for both regular and irregular MSB, but we do not find statistically significant differences between the two groups. Second, Yu et al. (2021) investigate possible explanations for MSB and find that when participants are nudged into reconsidering their choices the proportion of MSB drops from 31% to 10% implying a relationship between low understanding and MSB.[24] Third, Agranov and Ortoleva (2025) find no systematic relation between the extent to which decision-makers randomize and measures of cognitive abilities. Finally, in our experiment, we find no correlation between violations of FOSD and cognitive abilities. This result is surprising and, when paired with evidence from MPLs, suggests that this relation may

[24] Interestingly, note that this approach follows Gilboa (2009) definition of rationality as "feeling of embarrassment" when confronted with a previous choice, and it is in line with the recent approach adopted by Nielsen and Rehbeck (2022).

depend on features of the experimental design (i.e., violations of transitivity are easily identifiable by multiple-switching behaviour) as well as on the lotteries under study.

We believe that further research is needed to investigate the interaction between **PrAx** and **ConAx** within MPLs and the effect on the correlation with cognitive abilities as extensive literature has drawn conclusions that may potentially be misleading.

7.3 Discussion: model misspecification and our notion of consistency

Recognizing the importance of the underlying model of behaviour is crucial to interpreting the correlation between consistency and cognitive abilities. For example, if participants were explicitly instructed to choose according to a model, i.e., to maximize utility, departures from the model's predictions (WARP) would likely correlate with cognitive abilities, as they would reflect the ability to perform the task effectively. In practice, however, the choice of the model is endogenous and affects the observed correlation. For instance, heuristics may be preferred for their simplicity (e.g., thinking aversion) while randomization may be computationally burdensome as the problem becomes non-separable. As analysts, we do not know the correct model specification and often interpret the data, leaving the model implicit. Many studies have concluded, upon observing a positive correlation between consistency and cognitive abilities, that (most) inconsistencies were errors, implicitly assuming an ARUM specification. In this paper, we explicitly model inconsistencies as errors in a utility maximization problem and demonstrate that this model is misspecified for a subset of our participants. As a result, we uncover the underlying mechanisms behind the lack of correlation between consistency with WARP and cognitive abilities observed after controlling for confounding factors.

We acknowledge, however, that our measure of consistency also relies on model assumptions being a distance from utility maximization. This, on the one hand, allowed us to decouple utility maximization from cognitive sophistication. On the other hand, inconsistencies may hide regularities that our focus on WARP did not reveal. For example, weaker notions may be necessary conditions for high decision ability, as sophisticated decision-makers may explore reasonable departures from utility maximization. In the online appendix, we explore two of such weaker notions: "No Binary Cycle" and "Always Chosen" which are jointly equivalent to WARP (Manzini and Mariotti, 2007, Proposition 1). These properties interestingly categorize violations of WARP into two types: intransitivities in binary menus and menu effects, which may correlate differently with cognitive abilities. Notably, our results hold for both properties. Another example is conscious randomization which is characterized by a consistency property over mixtures of lotteries (Cerreia-Vioglio et al. 2019, see Appendix A), whereas our definition of WARP is on lotteries alone. It is possible that "consistent" randomizers have better decision-making ability than inconsistent ones but our design is not equipped to investigate this question. We believe this could be an interesting avenue for future work. To conclude, although we have shown that WARP is not a necessary condition for high decision-making ability, we cannot rule out that other notions of consistency may be.

8 Conclusion

Existing literature has documented a positive correlation between cognitive sophistication and consistent behaviour, and has consequently interpreted inconsistencies as errors. We have shown that this correlation could be driven by confounding factors, and isolating violations of consistency, we found no correlation. Heuristics and randomization are key mechanisms underlying our result. The former make non-sophisticated decision-makers consistent, while the latter make sophisticated ones appear inconsistent. Overall, this suggests that consistency with WARP is neither sufficient nor necessary for high decision-making ability, challenging the use of the language that, in economics, equates rationality and consistency.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00355-026-01656-8.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability Replication Package and data available here: https://github.com/DCaliari/ConsistencyRationality.git

Declaration

Conflict of interest None.

Appendix A: Randomization

Conscious randomization has been modelled by Machina (1985) and more recently by Cerreia-Vioglio et al. (2019), and denoted "deliberate" randomization. Decision-makers have a deterministic preference over the convex hull of a set of lotteries and randomize to obtain the optimum. We use the framework of Cerreia-Vioglio et al. (2019) to formalize this idea. Let $[w, b] \subseteq \Re$ be an interval on monetary prizes and Δ be the set of lotteries over $[w, b]$. Denote $\mathcal{A}$ as the collection of all finite, non-empty subsets of Δ; and $co(A)$ as the convex hull of $A \in \mathcal{A}$. A stochastic choice function ρ is a map that assigns to each A a probability distribution $\rho(A)$. Particularly, $\rho(A)$ is a compound lottery and $\overline{\rho(A)}$ is the induced lottery over monetary outcomes:

$$\overline{\rho(A)} = \sum_{q \in A} \rho(A)(q)q.$$

A stochastic choice function ρ has a Deliberate Stochastic Choice representation if there exists a complete preorder $\succeq$ over Δ such that:

1. For every $A \in \mathcal{A}$:
$$\overline{\rho(A)} \succeq q \quad \forall \quad q \in co(A)$$
2. For all $p, q \in \Delta$, $p >_{FOSD} q$ implies $p \succ q$.

Appendix B: Robustness checks: identification of the heuristics

B.1 Questionnaire

At the end of the experiment, we collected information through a questionnaire. We used open questions to ask which model participants adopted and identify both heuristics and randomization. In the Online Appendix, we provide examples of the participants' answers. Below, we replicate our analysis on WARP violations and Raven scores (Figs. 8 and 9), while in the Online Appendix, we also replicate the results regarding response times (Figs. 4 and 5).

First, our reported measures are highly correlated with the estimated extreme preferences. In particular, the correlation is 0.61 (p-value = 0.000) for **patient** participants,

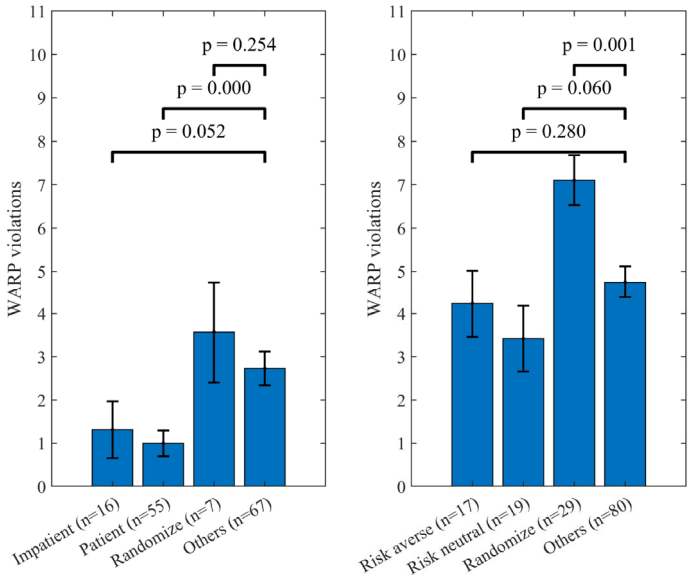

Fig. 12 WARP violations (heterogeneous analysis). Notes: The two plots report the average number of WARP violations for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests. Importantly, since our hypotheses in Sect. 5 are one-directional the statistical tests are one-tailed

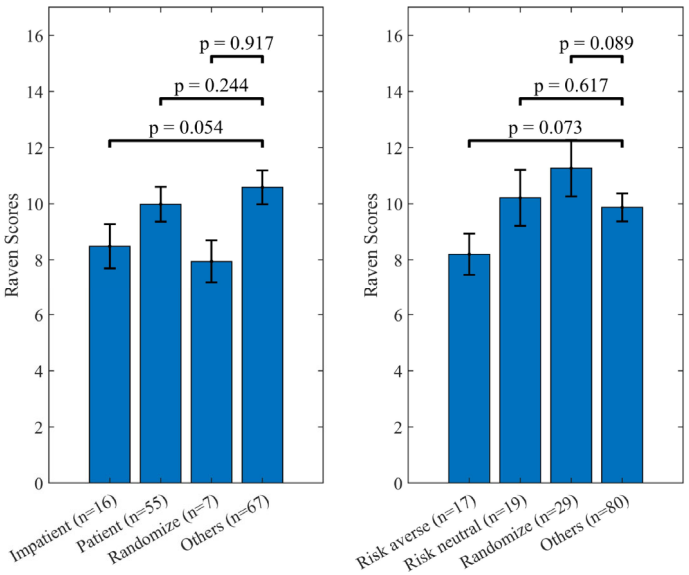

Fig. 13 Raven scores (heterogeneous analysis). Notes: The two plots report the average Raven scores for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests. Importantly, since our hypotheses in Sect. 5 are one-directional the statistical tests are one-tailed

0.58 (p-value = 0.000) for **impatient** ones, 0.41 (p-value = 0.000) for **risk-neutral** ones, and 0.29 (p-value = 0.000) for **risk-averse** ones. As the reader can see, comparing Sect. 5 with the above figures, the results are very similar both in Time and in Risk, with participants who reported the use of heuristics being more consistent and having no significant differences in decision-making ability (Figs. 12, 13).

B.2 Revealed preferences

As described in the Sect. 4.1, our structural approach is based on several assumptions. Here, we replicate our analysis on WARP violations and Raven scores (Figs. 8 and 9) using three (non-parametric) approaches to revealed preferences (similar to the previous section, other results are relegated to an online appendix): (i) minimum swaps (Apesteguia and Ballester 2015), (ii) sequential method (Horan and Sprumont 2016), and (iii) the reported ordinal preferences. These approaches map participants' choices into ordinal preferences (possibly incomplete in the case of Apesteguia and Ballester 2015). Our definition of extreme preferences is restrictive. From Tables 1 and 2, denote the delayed payment plans: One Shot Payment [OS], Decreasing [D], Constant [K], Increasing [I]; and the lotteries: Degenerate [D], Safe [S], Fifty-Fifty [50] and Risky [R]. **Impatient** and **patient** participants have preferences OS $\succ$ D $\succ$ K $\succ$ I and I $\succ$ K $\succ$ D $\succ$ OS, respectively. **Risk-averse** and **risk-neutral** participants have preferences D $\succ$ S $\succ$ 50 $\succ$ R and R $\succ$ 50 $\succ$ S $\succ$ D, respectively. The results are robust to all these identification approaches (Figs. 14, 15, 16, 17, 18, 19).

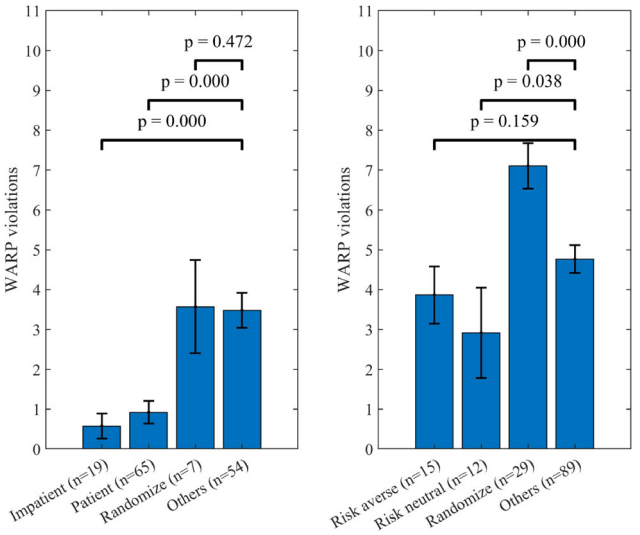

Fig. 14 WARP violations (heterogeneous analysis) with "minimum swaps" revealed preferences. Notes: The two plots report the average number of WARP violations for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests. Importantly, since our hypotheses in Sect. 5 are one-directional the statistical tests are one-tailed

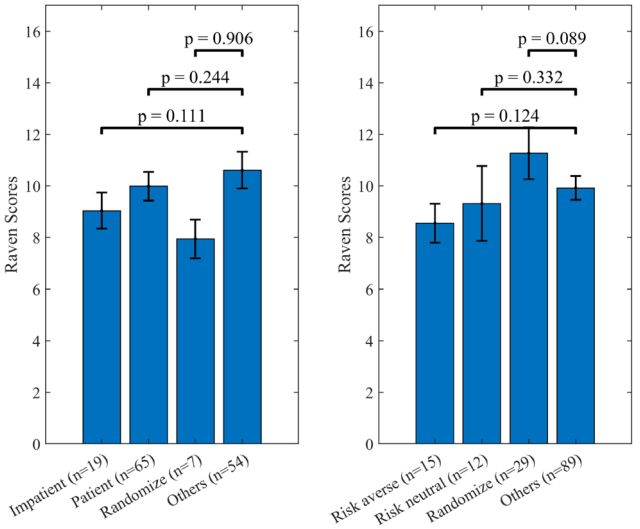

Fig. 15 Raven scores (heterogeneous analysis) with "minimum swaps" revealed preferences. Notes: The two plots report the average Raven scores for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests. Importantly, since our hypotheses in Sect. 5 are one-directional the statistical tests are one-tailed

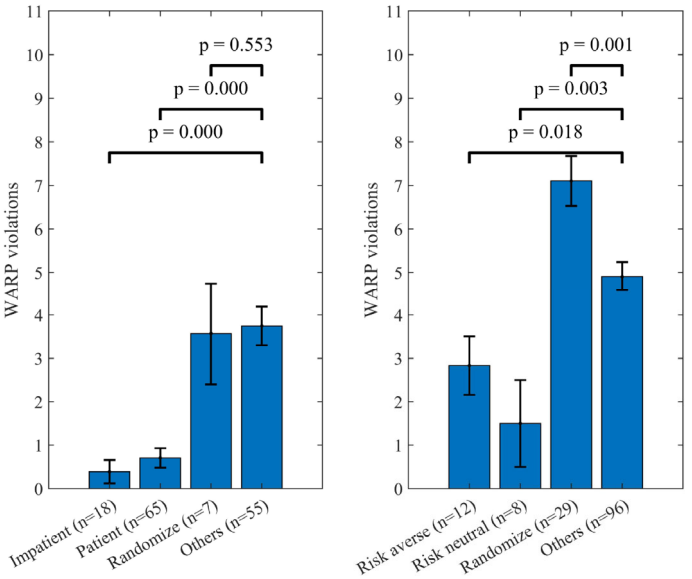

Fig. 16 WARP violations (heterogeneous analysis) with "sequential" revealed preferences. Notes: The two plots report the average number of WARP violations for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests. Importantly, since our hypotheses in Sect. 5 are one-directional the statistical tests are one-tailed

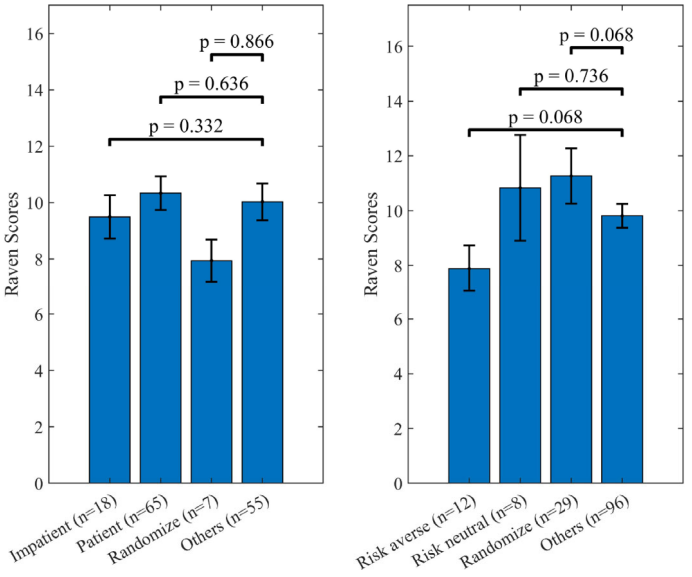

Fig. 17 Raven scores (heterogeneous analysis) with "sequential" revealed preferences. Notes: The two plots report the average Raven scores for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests. Importantly, since our hypotheses in Sect. 5 are one-directional the statistical tests are one-tailed

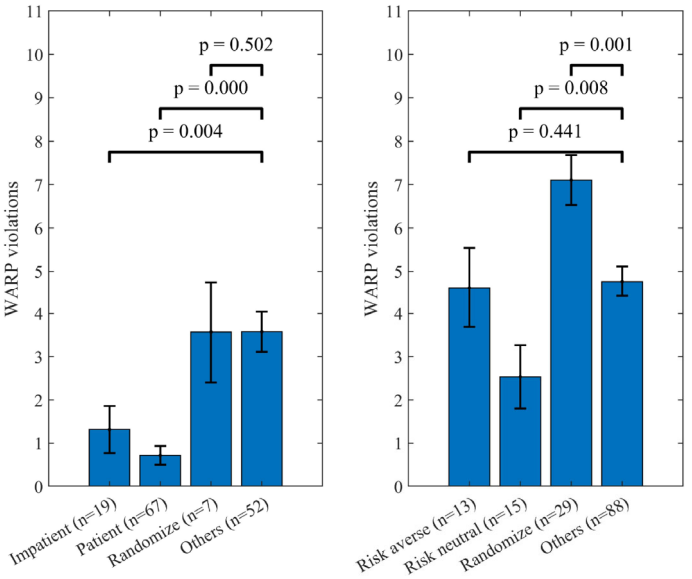

Fig. 18 WARP violations (heterogeneous analysis) with reported preferences. Notes: The two plots report the average number of WARP violations for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests. Importantly, since our hypotheses in Sect. 5 are one-directional the statistical tests are one-tailed

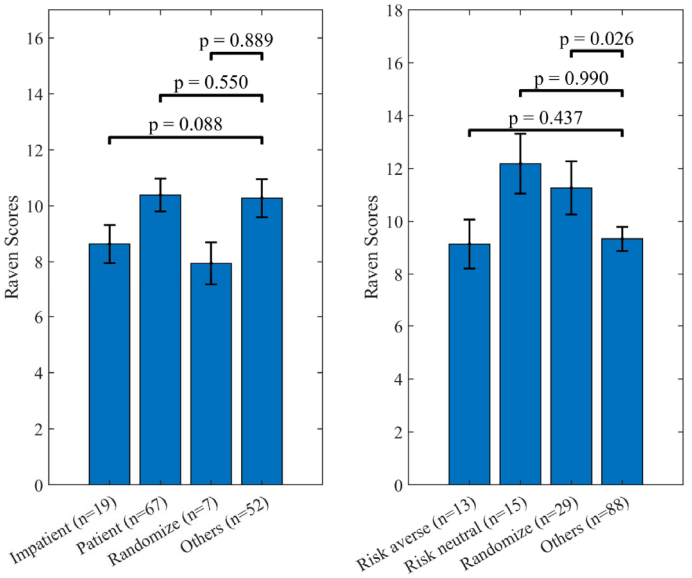

Fig. 19 Raven scores (heterogeneous analysis) with reported preferences. Notes: The two plots report the average Raven scores for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). On the x-axis, we include in parentheses the numerosity of each group. Finally, above the bars, we include the p-values from Welch's t-tests. Importantly, since our hypotheses in Sect. 5 are one-directional the statistical tests are one-tailed

Appendix C: Structural estimation

Complementing Sect. 4.1, we provide more details on our structural estimation. Let X be a set of alternatives and $\{U_\theta\}_{\theta\in\Omega}$ be a collection of utility functions characterized by the preference parameter θ. We focus on two functional forms of the utility functions: the CRRA utility function with preference parameter $\rho \in \mathbb{R}^+$ to model risk preferences[25] and the exponential discounting with preference parameter $\beta \in [0, 1]$ to model time preferences.

$$U_\rho^{CRRA}(x_1, x_2; p_1, p_2) = \sum_{i=1}^{2} p_i \frac{x_i^{1-\rho}}{1-\rho}$$

$$U_\beta^{EXP}(m_1, \ldots, m_5; t_1, \ldots, t_5) = \sum_{i=1}^{5} \beta^{t_i} m_i.$$

We assume the decision-maker can make a stochastic error modelled using a logit function with precision parameter λ. We use the "contextual approach" as described by Wilcox (2011) to limit the non-monotonic behaviour of the implied probabilities in the preference parameters. Namely, for each parameter θ the utilities of each alternative x is normalized as follows:

$$\hat{U}_\theta(x) = U_\theta(x) = \frac{U_\theta(x) - \min_{z\in X} U_\theta(z)}{\max_{z\in X} U_\theta(z) - \min_{z\in X} U_\theta(z)}$$

The decision-maker faces 11 choice menus (the set of all non-empty subsets of X) A_i indexed by $i = 1, \ldots, 11$. We denote (a, A_i) the observation in which the decision maker chooses a from the set A_i. In each menu A_i, the probability of selecting an alternative a, $Pr(a, A_i)$ is:

$$Pr(a, A_i) = \frac{e^{\lambda \hat{U}(a)}}{\sum_{b\in A_i} e^{\lambda \hat{U}(b)}}.$$

Estimation of individual preference and cognitive parameters Due to our interest in heterogeneous behaviour we use a latent-class model approach (Train 2008; Dardanoni et al. 2023). Let $s = 1, \ldots, S$ be the participants. A type $\omega = (\theta, \lambda)$ is a tuple of preference and precision parameter. The individual types $\omega_1, \ldots, \omega_S$ are generated by an unknown density $g(\omega)$. Each participant faces each of the 11 choice menus independently and each realisation of the unobserved parameter ω_s yields, independently for each participant $s = 1, \ldots, S$, observed choices $(a, A_i)_s$, realized according to the probabilities defined in the previous section. The observed choices of each participant s are summarized in the choice function c_s.

[25] The definition of CRRA utility function in the main text is not defined for $\rho = 1$, but it well-known that in such case $U_\rho^{CRRA}(x_1, \ldots, x_n; p_1, \ldots, p_n) = \sum_{i=1}^{n} p_i \log(x_i)$.

We estimate the population density $g(\omega)$ by maximum likelihood, and given the actual individual choices, we adopt the standard approach of using Bayes' Theorem to estimate the posterior distribution of parameters ω_s. Following Train (2008), we define a finite number of unobserved types in the population by appropriately discretizing θ in a finite grid of Θ values, and λ in a finite grid of Λ values. There are $\Theta \cdot \Lambda = \Omega$ unobserved types in the population, with an unknown distribution described by a $\Omega \times 1$ probability vector π. Notice that this formulation leaves the joint distribution of the two underlying deep parameters unrestricted in line with our interest in heterogeneity and can be viewed as a discrete non-parametric approximation of their joint distribution.

An estimator of π, i.e. the (discretized) population density of ω, can be obtained as follows. First, for each discretized value of ω, we can write the probability for each possible choice vector c_s. Let P denote the $S \times \Omega$ matrix arraying these probabilities for each participant and each discretized value of ω, that is $P_{s,j} = Pr(c_s \mid \omega_j)$.

Bayes rule implies that for each participant s, the $\Omega \times 1$ vector of posterior probabilities

$$Pr(\omega_j \mid c_s) = \frac{diag(p_s)\pi}{p_s \pi}$$

where p_s denotes the s row of P. Let H denote the $S \times \Omega$ matrix collecting the posterior probabilities for all S participants. Notice that, given the model and the discretized parameters, H is a function of π only. The following iterative procedure (EM-algorithm)[26] gives our estimator $\hat{\pi}$. Start with a value π_1 (typically uniformly distributed). At each iteration t: (1) compute $H(\pi_t)$; (2) compute $\pi'_{t+1} = 1'H(\pi_t)/S$, where 1 denotes the $S \times 1$ vector of ones; (3) Iterate until convergence.

From $\hat{\pi}$, we can use Bayes theorem as in the above formula to obtain the individual posterior distributions of the preference and precision parameters (θ_s, λ_s) from which we have plotted the expected values in Figs. 2 and 3.

C.1 Robustness check: utility specifications

In Sect. 4.1, we estimated preference and precision parameters using Exponential discounting in Time and CRRA utility in Risk. Our subsequent claims on response times and violations of IMP and SOSD rely on a correct cardinal specification of the functionals. Here, we provide a robustness check by replicating the structural estimation with other well-known cardinal definitions. In Time, beyond exponential discounting where the discount factor follows the power function $\beta^t = \frac{1}{(1+\delta)^t}$, we focus on hyperbolic discounting, i.e. $\beta(t) = \frac{1}{1+\delta t}$, and quasi-hyperbolic discounting, i.e. $\beta(0) = 1$ and $\beta(t) = \gamma \frac{1}{(1+\delta)^t}$ for all $t \geq 1$, with $\gamma \in (0, 1]$. In Risk, beyond the CRRA utility function, we focus on three other specifications: the CARA utility function, i.e. $\frac{(1-e^{-\rho x})}{\rho}$, and the certainty equivalents of both CRRA and CARA utility functions. Below, in Tables 4 and 5, we report the correlations between the estimates of δ and ρ in the different specifications, as well as the precision parameters. The correlations between the preference parameters are close to one in Time, while slightly

[26] Note that we keep fixed the discretized values of the parameters. The same estimation can be performed maximizing the values of the parameters as described in Arcidiacono and Jones (2003). This extra step is computationally burdensome, does not change the results, and is beyond the scope of the paper.

Table 4 Correlations between preference parameter estimates

	Exp	Hyp	Q-Hyp	
Exp	1.000	0.998	0.993	
Hyp	0.998	1.000	0.989	
Q-Hyp	0.993	0.989	1.000	
	CRRA	CARA	CRRA/ce	CARA/ce
CRRA	1.000	0.977	0.892	0.896
CARA	0.977	1.000	0.948	0.950
CRRA/ce	0.892	0.950	1.000	0.995
CARA/ce	0.896	0.950	0.995	1.000

Notes: We report the linear correlations between the individual preference parameter estimates under different utility specifications

Table 5 Correlations between precision parameter estimates

	Exp	Hyp	Q-Hyp	
Exp	1.000	0.874	0.978	
Hyp	0.874	1.000	0.826	
Q-Hyp	0.978	0.826	1.000	
	CARA	CARA	CRRA/ce	CARA/ce
CRRA	1.000	0.967	0.892	0.935
CARA	0.967	1.000	0.933	0.978
CRRA/ce	0.892	0.933	1.000	0.983
CARA/ce	0.935	0.978	0.983	1.000

Notes: We report the linear correlations between the individual precision parameter estimates under different utility specifications

lower in Risk, as expected given the discussion of the differences between the two environments in Apesteguia and Ballester (2018). The precision parameters are also highly correlated. This evidence supports the robustness of our approach. In the Online Appendix, we replicate all main figures under all these specifications.

C.2 Validation

Finally, we provide an additional sanity check of our structural estimates using the three approaches introduced in Appendix B.2. Figures 20 and 21 show strong correlations between the (revealed) preferences from all of these methodologies and our estimated parameters.

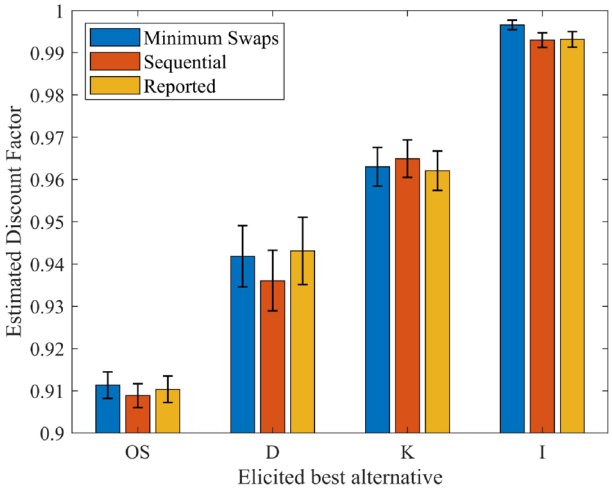

Fig. 20 Discount Factor and elicited preferences. Notes: This figure and Fig. 21 report on the y-axis the average preference parameters. On the x-axis, there is the best element either revealed with the minimum swaps and the sequential algorithms, or directly reported

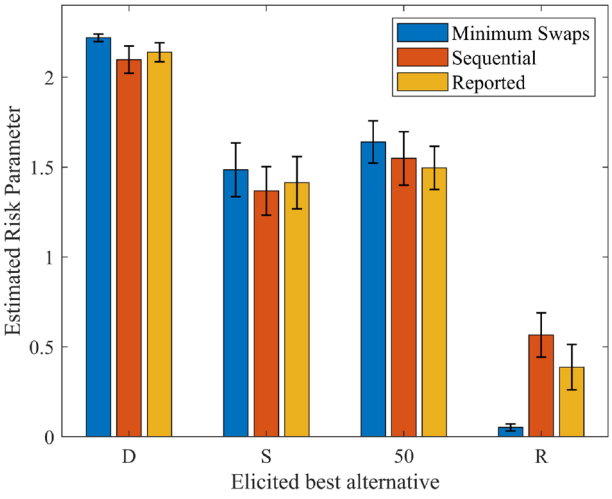

Fig. 21 CRRA risk parameter and elicited preferences. Notes: Figure 20 and this figure report on the y-axis the average preference parameters. On the x-axis, there is the best element either revealed with the minimum swaps and the sequential algorithms, or directly reported

References

Afriat SN (1967) The construction of utility functions from expenditure data. Int Econ Rev 8(1):67–77

Agranov M, Ortoleva P (2017) Stochastic choice and preferences for randomization. J Polit Econ 125(1):40–68

Agranov M, Ortoleva P (2025) Ranges of randomization. Rev Econ Stat 107(6):1702–1713

Agranov M, Schotter A, Trevino I (2025) Complex for whom? An experimental approach to subjective complexity. Working paper

Alós-Ferrer C, Fehr E, Netzer N (2021) Time will tell: recovering preferences when choices are noisy. J Polit Econ 129(6):1828–1877

Andersen S, Harrison GW, Lau MI, Rutstrom EE (2008) Eliciting time and risk preferences. Econometrica 76(3):583–618

Andersson O, Holm HJ, Tyran J-R, Wengstrom E (2016) Risk aversion relates to cognitive ability: preferences or noise? J Eur Econ Assoc 14(5):1129–1154

Apesteguia J, Ballester MA (2015) A measure of rationality and welfare. J Polit Econ 6(123):1278–1310

Apesteguia J, Ballester MA (2018) Monotone stochastic choice models: the case of risk and time preferences. J Polit Econ 126(1):74–106

Arcidiacono P, Jones JB (2003) Finite mixture distributions, sequential likelihood and the EM algorithm. Econometrica 71(3):933–946

Arrow KJ (1959) Rational choice functions and orderings. Economica 26(102):121–127

Ballinger TP, Wilcox NT (1997) Decisions, error and heterogeneity. Econ J 107(443):1090–1105

Banks J, Carvalho LS, Perez-Arce F (2019) Education, decision making, and economic rationality. Rev Econ Stat 101(3):428–441

Block H, Marschak J (1960) Random orderings and stochastic theories of responses. In: Contributions to probability and statistics. Stanford University Press

Burks SV, Carpenter JP, Goette L, Rustichini A (2009) Cognitive skills affect economic preferences, strategic behavior, and job attachment. Proc Natl Acad Sci 106(19):7745–7750

Carvalho L, Silverman D (2019) Complexity and sophistication. NBER Working Paper 26036

Cerreia-Vioglio S, Dillenberger D, Ortoleva P, Riella G (2019) Deliberately stochastic. Am Econ Rev 7(109):2425–2445

Chew SH, Miao B, Shen Q, Zhong S (2022) Multiple-switching behavior in choice-list elicitation of risk preference. J Econ Theory 204:105510

Choi S, Fisman R, Gale D, Kariv S (2007) Consistency and heterogeneity of individual behavior under uncertainty. Am Econ Rev 97(5):1921–1938

Choi S, Kariv S, Müller W, Silverman D (2014) Who is (more) rational? Am Econ Rev 104(6):1518–50

Dardanoni V, Manzini P, Mariotti M, Petri H, Tyson CJ (2023) Mixture choice data: revealing preferences and cognition. J Polit Econ 131(3):687–715

Dembo A, Kariv S, Polisson M, Quah JK (2022) Ever since allais. Working paper

Filippin A, Crosetto P (2016) A reconsideration of gender differences in risk attitudes. Manag Sci 62(11):3138–3160

Fischbacher U (2007) z-tree: Zurich toolbox for ready-made economic experiments. Exp Econ 10:171–178

Fishburn PC, Rubinstein A (1982) Time preference. Int Econ Rev 23(3):677–694

Gilboa I (2009) Theory of decision under uncertainty, vol 45. Cambridge University Press, Cambridge

Hansen L, Holt C, Peled D (1978) A note on first degree stochastic dominance. Econ Lett 1(4):315–319

Hey J (2001) Does repetition improve consistency? Exp Econ 4(1):5–54

Hey J, Carbone E (1995) Stochastic choice with deterministic preferences: an experimental investigation. Econ Lett 47(2):161–167

Holt CA, Laury SK (2002) Risk aversion and incentive effects. Am Econ Rev 92(5):1644–1655

Horan S, Sprumont Y (2016) Welfare criteria from choice: an axiomatic analysis. Games Econom Behav 99:56–70

Kariv S, Silverman D (2013) An old measure of decision-making quality sheds new light on paternalism. J Inst Theor Econ 169(1):29–44

Kreps DM (2015) Choice, dynamic choice, and behavioral economics. Stanford Graduate School of Business, lecture. http://stanford.io/1GxjZfg

Machina M (1985) Stochastic choice functions generated from deterministic preferences over lotteries. Econ J 95:575–594

Mahmoud O (2017) On the consistency of choice. Theor Decis 83:547–572

Manzini P, Mariotti M (2007) Sequentially rationalizable choice. Am Econ Rev 97(5):1824–1839
Manzini P, Mariotti M, Mittone L (2010) Choosing monetary sequences: theory and experimental evidence. Theor Decis 69(3):327–354
McCausland WJ, Davis-Stober C, Marley A, Park S, Brown N (2019) Testing the random utility hypothesis directly. Econ J 130:183–207
Nielsen K, Rehbeck J (2022) When choices are mistakes. Am Econ Rev 112(7):2237–2268
Oprea R (2024) Decisions under risk are decisions under complexity. Am Econ Rev 114(12):3789–3811
Rieskamp J, Busemeyer JR, Mellers BA (2006) Extending the bounds of rationality: evidence and theories of preferential choice. J Econ Lit 44(3):631–661
Rubinstein A (2013) Response time and decision making: an experimental study. Judgement Decis Mak 8(5):540–551
Sen A (1971) Choice functions and revealed preference. Rev Econ Stud 38(3):307–317
Sopher B, Narramore MJ (2000) Stochastic choice and consistency in decision making under risk: an experimental study. Theor Decis 48(4):323–350
Strzalecki T (2025) Stochastic choice theory. Econometric Society monograph. Cambridge University Press, Cambridge
Sugden R (1991) Rational choice: a survey of contributions from economics and philosophy. Econ J 101(407):751–785
Train KE (2008) EM algorithms for nonparametric estimation of mixing distributions. J Choice Model 1(1):40–69
Tversky A, Russo JE (1969) Substitutability and similarity in binary choices. J Math Psychol 6(1):1–12
Wilcox NT (2011) 'Stochastically more risk averse:' a contextual theory of stochastic discrete choice under risk. J Econometrics 162(1):89–104
Yu CW, Zhang YJ, Zuo SX (2021) Multiple switching and data quality in the multiple price list. Rev Econ Stat 103(1):136–150