Online Appendix

ON THE RELATION BETWEEN RATIONALITY AND CONSISTENCY

Daniele Caliari*

January 19, 2026

Premise: information to the reader

This Online Appendix is organized as follows: section 1 provides an extensive description of the alternatives. Section 2 describes the questions and their orders. Section 3 contains additional analysis. In details, section 3.1 shows the remarkable difference in consistency between Time and Risk, section 3.2 replicates the figures in the paper using alternative cardinal utility functions; section 3.3 proposes an alternative unstructured cardinal approach, section 3.4 proposes an alternative ordinal approach, section 3.5 replicates the analysis using the Random Parameter Model [RPM] approach, section 3.6 proposes alternative definitions of extreme preferences, section 3.7 replicates the analysis using non-parametric methods of revealed preferences, section 3.8 reports results on violations of Stationarity and Independence, section 3.9 reports robustness checks on the measurement of consistency; section 3.10 reports robustness checks on the measurerement of cognitive abilities; section 3.11 reports the replication of our test of ARUM on the restricted sample without participants who use heuristics or randomization. Section 4 presents the result of the questionnaire and some samples of open answers from the participants. Finally, Sections 5 and 6 present the instructions and main screens of the experiment: one screen describing a choice between lotteries, one describing a choice between delayed payment plans, and one describing the direct elicitation of preferences.

1 Alternatives

Following Agranov & Ortoleva (2017), and differently from Tversky & Russo (1969), Manzini & Mariotti (2010), McCausland et al. (2019), we construct the main alternatives with no clear domination. Each main alternative is matched to the dominated alternatives. We run a pilot experiment in the game theory classes at Queen Mary University of London to confirm the presence of heterogeneity in preferences.

*Berlin Social Science Center, WZB. E-mail: daniele.caliari@wzb.eu

ALTERNATIVES		MONTHS 0	3	6	9	12
ONE SHOT (OS)		**160**	**0**	**0**	**0**	**0**
	M(OS)	140	0	0	0	0
	Im(OS)	0	160	0	0	0
DECREASING (D)		**110**	**50**	**25**	**0**	**0**
	M(D)	100	40	10	0	0
	Im(D)	0	110	50	25	0
CONSTANT (K)		**50**	**50**	**50**	**50**	**0**
	M(K)	40	40	40	40	0
	Im(K)	0	50	50	50	50
INCREASING (I)		**0**	**15**	**40**	**170**	**0**
	M(I)	0	10	20	160	0
	Im(I)	0	0	15	40	170
NEUTRAL						
	Neu1	15	55	30	20	5
	Neu2	5	20	30	55	15

Notes: The amounts are described in Token with an exchange rate of 20:1 pounds. The alternatives "M" are monotonically dominated; "Im" are impatience-dominated; "Neutral" are confounding alternatives.

Table 1: List of Delayed Payment Plans

ALTERNATIVES		TOKEN		PROBABILITIES		EV	SD
DEGENERATE (D)		**50**	**0**	**1**	**0**	50	0
	F1(D)	45	0	1	0	45	0
	F2(D)	50	0	0.95	0.05	47.5	10.9
	So(D)	51	0	0.95	0.05	48.45	11.12
SAFE (S)		**65**	**25**	**0.8**	**0.2**	57	16
	F1(S)	60	20	0.8	0.2	52	16
	F2(S)	65	25	0.75	0.25	55	17.32
	So(S)	70	15	0.7	0.3	53.5	25.2
FIFTY-FIFTY (50)		**95**	**20**	**0.5**	**0.5**	57.5	37.5
	Sim	100	0	0.5	0.5	50	50
	F2(50)	95	20	0.45	0.55	53.75	37.31
	So(50)	110	20	0.4	0.6	56	44.09
RISKY (R)		**300**	**5**	**0.2**	**0.8**	64	118
	F1(R)	250	5	0.2	0.8	54	98
	F2(R)	300	5	0.15	0.85	49.25	105.34
	So(R)	500	5	0.1	0.9	54.5	148.5

Notes: The amounts are described in Token with an exchange rate of 10:1 pounds. The alternatives "F1" and "F2" are first-order stochastically dominated; "So" are second-order stochastically dominated; "Sim" is a simple confounding alternative which was created to check for "simplicity seeking" heuristics.

Table 2: List of Lotteries

2 Order of Questions

The construction of the questions is based on the literature on the order effect and the presence of different behavioural effects that could induce violations of preference axioms such as asymmetric

dominance or choice overload. A survey can be found in Day et al. (1987). They identify four factors as different sources of order effects and denote them as:

1. **Discovered preference hypothesis** (Plott, 1993) and **Institutional learning**. The first refers to respondents that, when faced with new decisions in unfamiliar environments, exhibit significant randomness in initial decisions. The second is related to the fact that respondents may have never experienced a lab experiment and surely they have never experienced the present design;

2. **Fatigue**: respondents may get tired of the choice task, especially if it is repeated many times. Hence they could exhibit higher randomness in later tasks;

3. **Starting point effect**: Respondents may create a reference point in some choice task and base subsequent choices on that.

QUESTIONS		ALTERNATIVES											
1		OS	D										
2		OS	K										
3		OS	I										
4		D	K										
5	M	D	I										
6	A	K	I										
7	I	OS	D	K									
8	N	OS	D	I									
9		OS	K	I									
10		D	K	I									
11		OS	D	K	I								
12		D	I	M(D)									
13	A	D	I	M(I)									
14	D	D	I	Im(D)									
15		D	I	Im(I)									
16		Im(OS)	Im(D)	Im(K)	Im(I)								
17		M(OS)	M(D)	M(K)	M(I)								
18		Neu1	Neu2										
19		M(OS)	M(D)	Neu1	Neu2								
20		Im(K)	Im(I)	Neu1	Neu2								
21		D	I	M(D)	M(I)	Im(D)	Im(I)	Neu1	Neu2				
22	B	OS	D	K	I	M(OS)	M(D)	M(K)	M(I)	Neu1	Neu2		
23	I	OS	D	K	I	Im(OS)	Im(D)	Im(K)	Im(I)	Neu1	Neu2		
24	G	D	I	M(OS)	M(K)	Im(OS)	Im(K)	Neu1	Neu2				
25		OS	D	K	I	M(OS)	M(D)	M(K)	M(I)	Im(OS)	Im(D)	Im(K)	Im(I)

Table 3: List of questions in Time

QUESTIONS		ALTERNATIVES									
1	MAIN	D	S								
2		D	50								
3		D	R								
4		S	50								
5		S	R								
6		50	R								
7		D	S	50							
8		D	S	R							
9		D	50	R							
10		S	50	R							
11		D	S	50	R						
12	AD	S	R	F1(S)							
13		S	R	So(S)							
14		S	R	F1(R)							
15		S	R	So(R)							
16		F2(D)	F2(S)	So(50)							
17		F2(D)	F2(R)	So(50)							
18		F2(S)	F2(50)	F2(R)							
19		F2(50)	Sim								
20	BIG	D	S	50	R	So(D)	So(S)	So(50)	So(R)	F1(D)	F1(50)
21		D	S	50	R	So(D)	So(S)	So(50)	So(R)		
22		D	S	50	R	F2(D)	F2(S)	F2(50)	F2(R)		
23		Sim	S	50	R	So(D)	So(S)	So(50)	So(R)		
24		Sim	R	S	50	D	So(D)	So(S)	So(R)		
25		D	S	50	R	F1(D)	F1(S)	F1(50)	F1(R)	Sim	F2(R)

Table 4: List of questions in Risk

In order to reduce the magnitude of the first two effects we construct confounding alternatives, as described in Section 1. We also avoid participants having immediate knowledge of the MAIN alternatives by designing a heterogeneous list of questions. This design feature has been previously exploited by Badescu & Weiss (1987), and suggested by Charness et al. (2012). As described in Tables 3 and 4, the questions are divided into three main domains: MAIN, AD and BIG; the remaining questions are neutral questions.

As suggested by Ladenburg & Olsen (2008) and Carlsson et al. (2012), we include three example questions before each part of the experiment in order to reduce institutional learning. These questions have irrelevant alternatives and are mixed in the number of alternatives so as not to affect the preferences or create reference points.

Fatigue is not a major issue. Numerous studies find no or small differences in preferences due to fatigue in a sequence of identical choice problems (Carlsson et al., 2012) (note that our choice problems are not identical). We set the number of questions to 50, more or less in line with the literature: Agranov & Ortoleva (2017) asked 70 questions, Cavagnaro & Davis-Stober (2014) 120 questions, Manzini & Mariotti (2010) 22 questions, Barberá & Neme (2017) 16 questions. However, the first two experiments involved pairs of lotteries and so they have a simpler design compared to our experiment; while the second two experiments lasted only 15 to 30 minutes.

TIME		RISK	
ORDER 1	ORDER 2	ORDER 1	ORDER 2
21	24	23	21
14	14	13	13
1	4	1	4
23	16	20	24
10	3	10	3
13	13	14	14
3	1	3	1
19	17	16	18
7	8	7	8
22	22	22	22
2	10	2	10
18	18	19	19
11	11	11	11
20	19	17	16
15	15	15	15
9	9	9	9
25	25	25	25
8	6	8	6
17	23	18	20
4	2	4	2
12	12	12	12
16	7	24	7
6	20	6	17
5	5	5	5
24	21	21	23

Table 5: The four orders of questions

In Table 5 we present the four orders that have been used. The construction, both in Time and Risk, followed a structural randomization. Namely, we divided the alternatives into the four similarity groups shown in Table 3 and 4. Then, we randomized the positions in order to avoid, as much as possible, similar questions arising consecutively.

3 Additional analysis and robustness checks

3.1 Consistency in Time and Risk

Participants' consistency differs significantly between Time and Risk. Figure 1 displays the number of violations of WARP in Risk on the x-axis and those in Time on the y-axis. The area of the circles is proportional to the number of participants. The correlation is low and mostly driven by a small portion of consistent participants in both Time and Risk. The magnitude of the WARP violations is similarly important. Figure 2 presents the distributions of violations in Time, Risk, and those that would have been made by a decision-maker who uniformly randomizes. A Kolmogorov-Smirnov test reports a highly significant difference between the three distributions Time/Risk (p-value < 0.01), Time/Random (p-value < 0.01), and Risk/Random (p-value < 0.01). To give an idea of the difference in magnitude, on average, participants display 1.96 WARP violations in Time and 4.99 in Risk (t-test, p-value < 0.01). This large difference is not surprising and replicates the findings of Agranov & Ortoleva (2017) who find that while 90% of the participants chose to randomize in Risk, only 34% did so in Time.

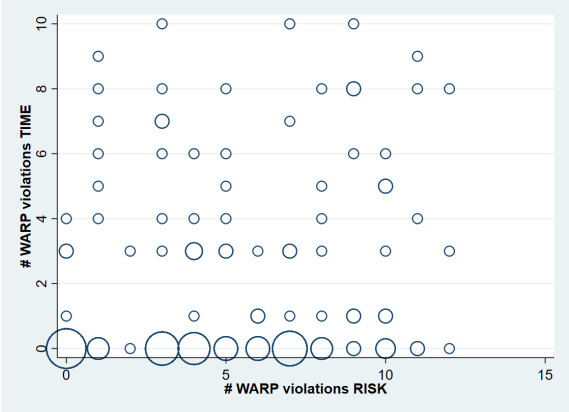


Figure 1: Violations of WARP between Time and Risk.

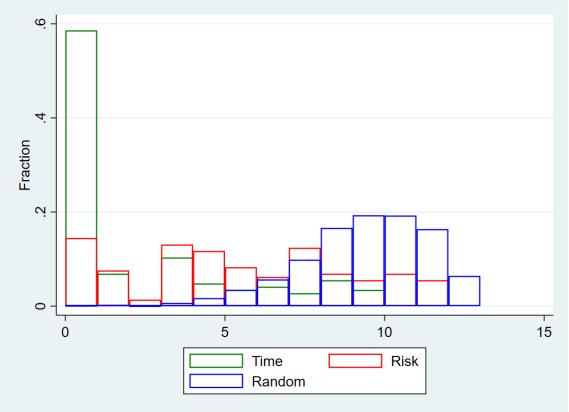


Figure 2: Distribution of the violations of WARP.

3.2 Robustness check I: alternative cardinal specifications of the utility function

Our main analysis relies on Exponential discounting in Time and CRRA utility in Risk. In the appendix, we replicated the analysis with alternative cardinal specifications, showing a strong correlation between the estimates of the preference and precision parameters. Here, we reproduce the paper's main figures. In Time, the two alternative specifications are the hyperbolic and the quasi-hyperbolic discounting models (from left to right in the figures). In Risk, the three alternative specifications are the CARA utility function and the certainty equivalents of CRRA and CARA utility functions (from left to right in the figures).

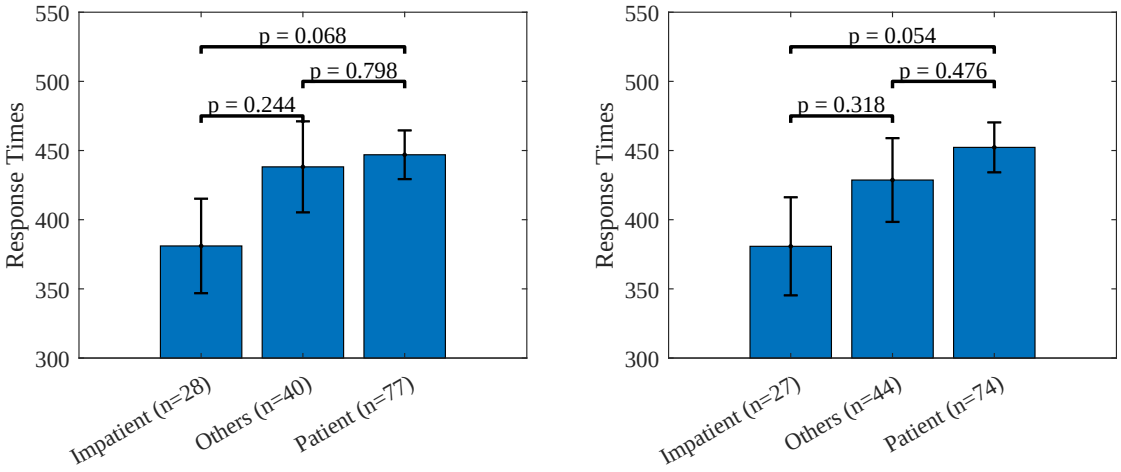


Figure 3: Response times in Time.

Notes: the bar plots report the sum of the response times in all questions, and the participants are grouped by their preferences. The left panel regards the hyperbolic discounting, the right panel the quasi-hyperbolic discounting. 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.

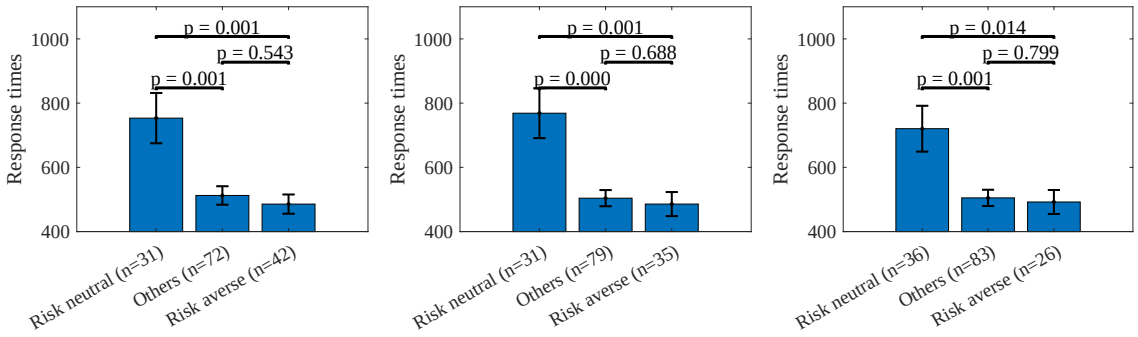


Figure 4: Response times in Risk.

Notes: the bar plots report the sum of the response times in all questions, and the participants are grouped by their preferences. The left panel regards the CARA utility, the middle panel the CRRA-certainty equivalent, and the right panel the CARA-certainty equivalent. 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.

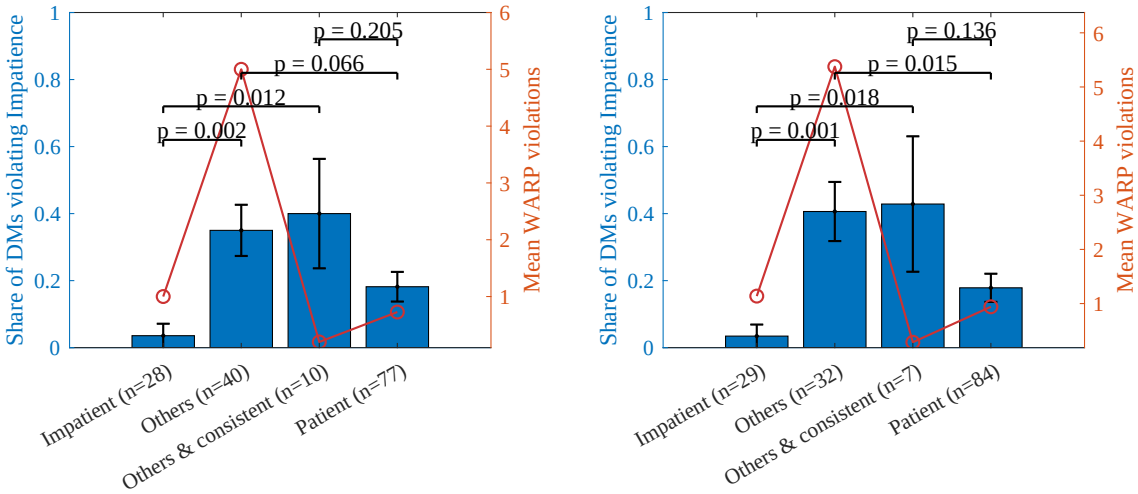


Figure 5: Share of participants violating IMP in Time.

Notes: the bar plots report the share of participants who violate IMP within each group defined by the preferences. The left panel regards the hyperbolic discounting, the right panel the quasi-hyperbolic discounting. The group "consistent" is defined as those participants with WARP violations lower or equal to 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.

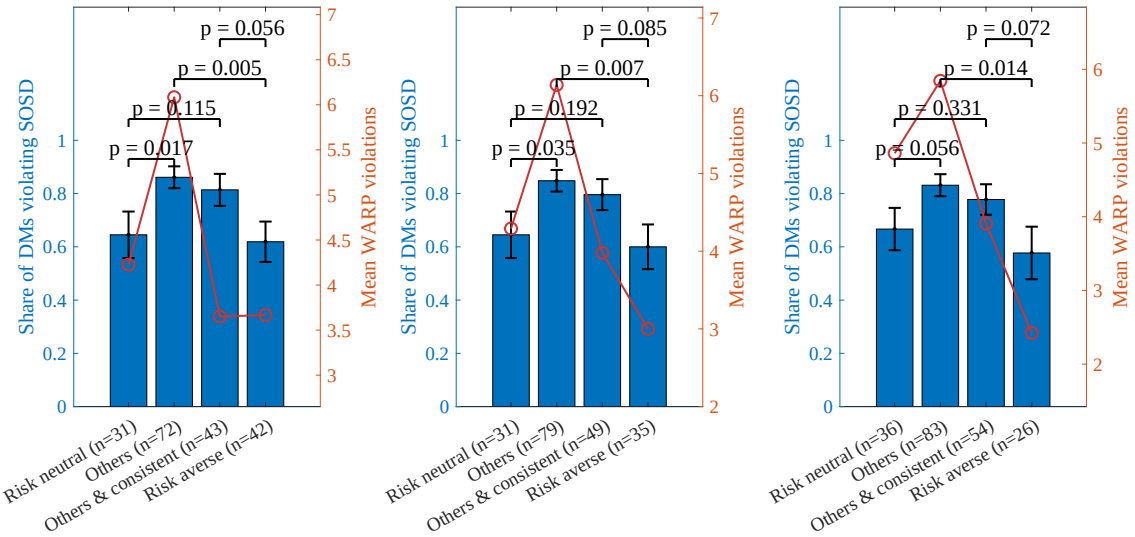


Figure 6: Share of participants violating SOSD in Risk.

Notes: the bar plots report the share of participants who violate SOSD within each group defined by the preferences. The left panel regards the CARA utility, the middle panel the CRRA-certainty equivalent, and the right panel the CARA-certainty equivalent. The group "consistent" is defined as those participants with WARP violations lower than or equal to 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.

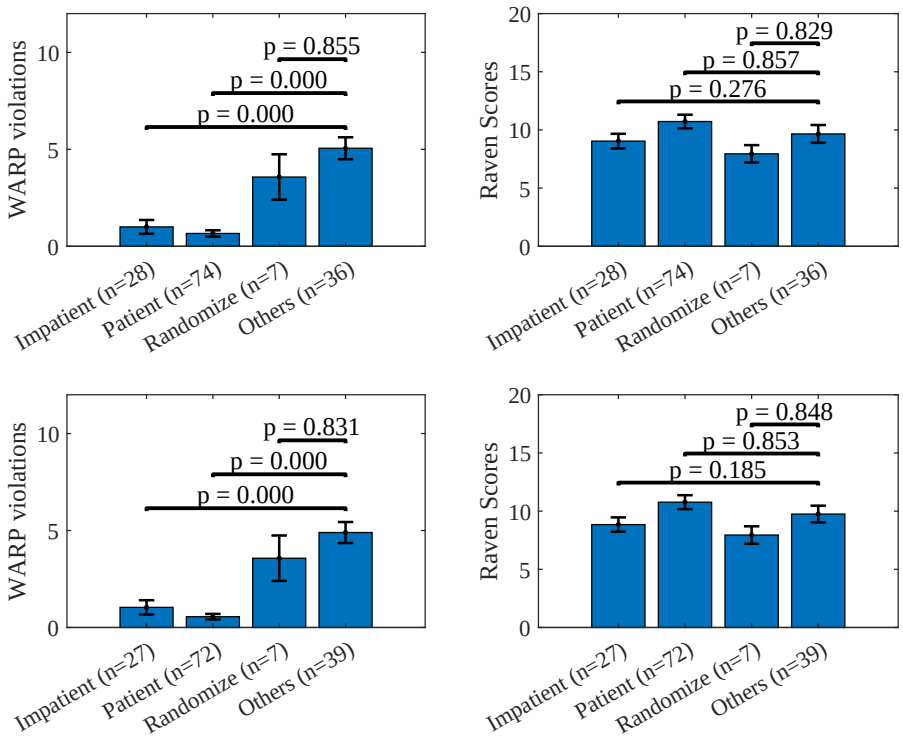


Figure 7: WARP violations and Raven scores in Time.

Notes: the plots on the left report the average number of WARP violations and the two on the right report the average Raven scores for for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). The top panels regard the hyperbolic discounting, the bottom panels the quasi-hyperbolic discounting. 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.

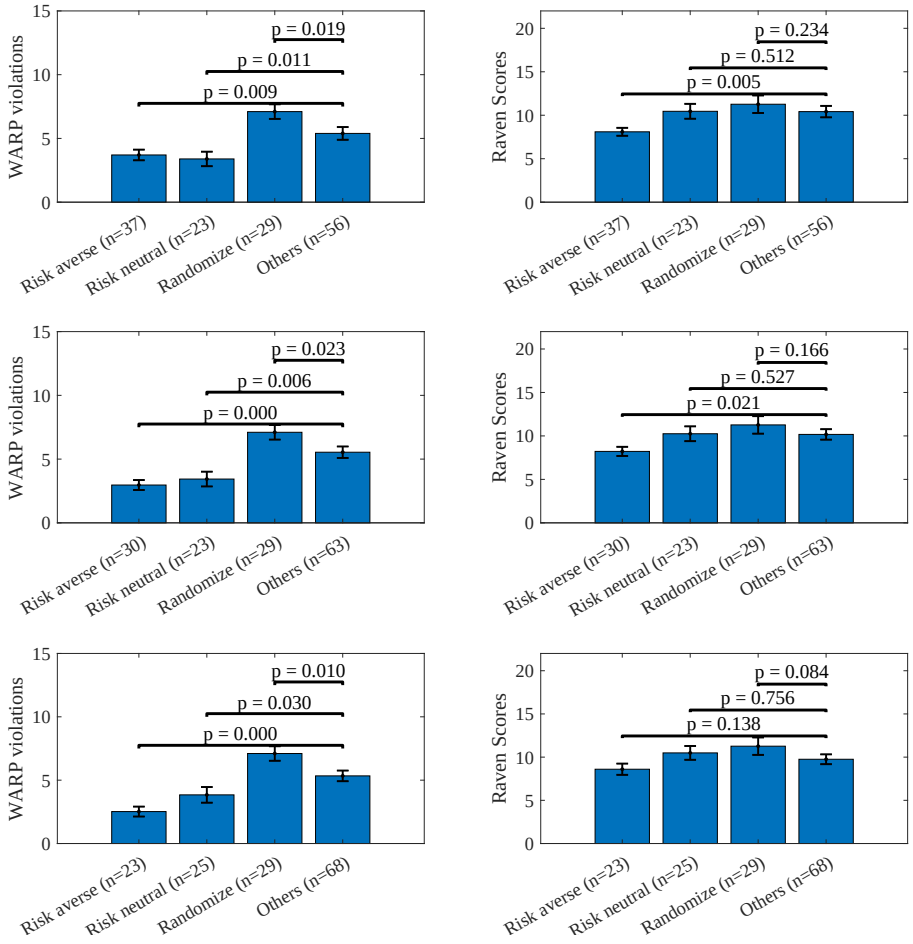


Figure 8: WARP violations and Raven scores in Risk.

Notes: the plots on the left report the average number of WARP violations and the two on the right report the average Raven scores for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). The top panels regard the CARA utility, the middle panels the CRRA-certainty equivalent, and the bottom panels the CARA-certainty equivalent. 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.

3.3 Robustness check II: (unstructured) Cardinal approach

In this section, we estimate preferences without making cardinal assumptions on the utility function. This can be achieved by estimating a logit model with a fixed precision parameter and varying utility differences. This contrasts with our approach in the paper, in which utility differences are "fixed" by our choice of the utility functional, whereas the precision parameter is estimated. Naturally, in this model, utility differences measure the subjects' choice variability, as explained below, while an estimate of their preferences can be obtained from the resulting ordinal ranking. Note that, since the utility function is unconstrained, ordinal rankings are also unconstrained, i.e., they are not restricted by a specific utility functional, and so this approach is similar to non-parametric revealed preference techniques such as the minimum swaps (Apesteguia & Ballester, 2015), and the sequential method (Horan & Sprumont, 2016) that we discuss in the paper.

To focus on utility differences, we estimate the following model. For all menus $A \subset \{x, y, z, w\}$ for each individual

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

under the restriction that $\sum_{b\in\{x,y,z,w\}} u(b)$ is constant. Note that this is a Fechnerian model as the probability depends on the utility differences (see e.g. Fudenberg et al. (2015) for a discussion).

We replicate our analysis in the paper by categorizing as extreme preferences the following utility functions: u(OS) > u(D) > u(K) > u(I) are **impatient**, u(I) > u(K) > u(D) > u(OS) are **patient**; and u(D) > u(S) > u(50) > u(R) are **risk-averse**, u(R) > u(50) > u(S) > u(D) are **risk-neutral**. This approach is equivalent to the one we used in Appendix B.2. We confirm that utility differences are a measure of consistency. The correlation between WARP violations and the individual mean utility difference, i.e. $\frac{1}{6}\sum_{a,b\in\{x,y,z,w\}} |u(a)-u(b)|$, is -0.88 and -0.86 in Time and Risk, respectively.

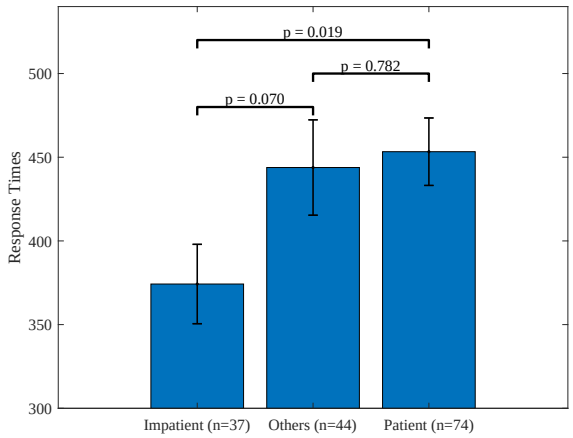


Figure 9: Response times in Time.

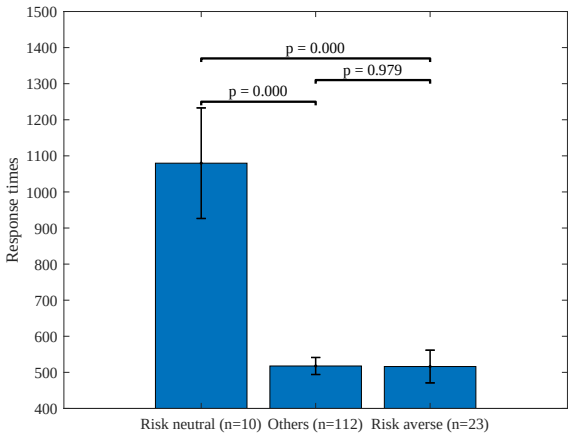


Figure 10: Response times in Risk.

Notes: the bar plots report the sum of the response times in all questions, and the participants are grouped by their preferences. 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.

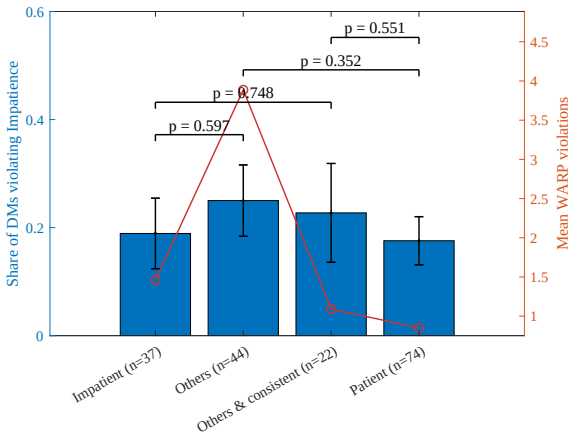


Figure 11: Share of participants violating IMP in Time.

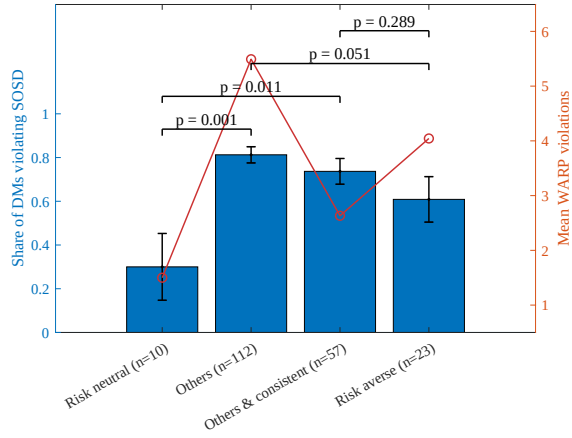
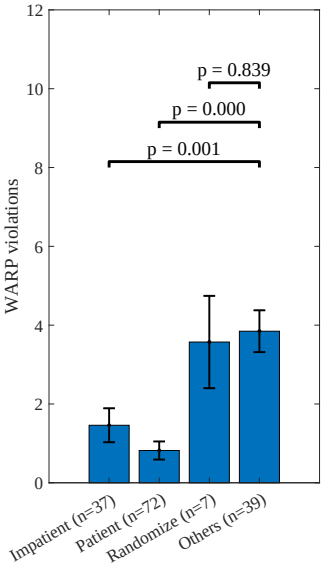


Figure 12: Share of participants violating SOSD in Risk.

Notes: the bar plots report the share of participants who violate IMP and SOSD within each group defined by the preferences. The group "consistent" is defined as those participants with WARP violations lower or equal to 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.



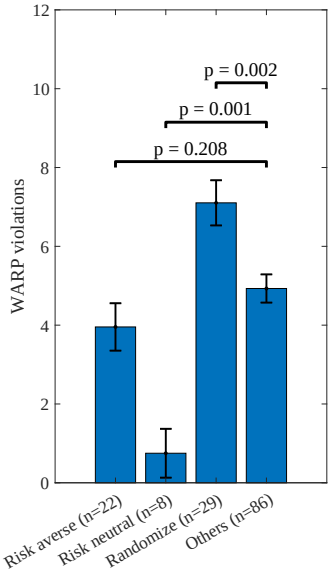
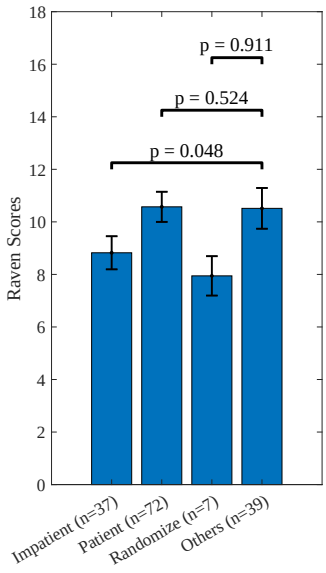


Figure 13: WARP violations: unstructured cardinal approach.



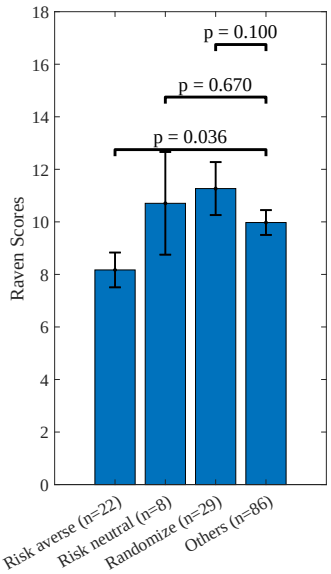


Figure 14: Raven scores: unstructured cardinal approach.

Notes: the two plots on the left report the average number of WARP violations and the two on the right 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.

3.4 Robustness check III: Ordinal approach

Another alternative approach to estimating preference parameters is the ordinal approach. We partition the preference parameter space of a specific utility functional with the associated ordinal preference as in Apesteguia et al. (2017). We denote each of these preferences P_θ for all parameters θ. We replicate our main analysis, focusing on exponential discounting and CRRA utility. For each individual, we estimate the probability distribution over the set of preferences, denoted $\hat{\mu}$, and obtain an estimate of the preference parameter by considering the median preference. To categorize individuals with extreme preferences, we consider those with $\hat{\mu}(P_{\bar{\theta}}), \hat{\mu}(P_{\underline{\theta}}) \geqslant 0.5$ where $\bar{\theta}, \underline{\theta}$ are the maximum and minimum parameters relevant for our partition.

Correlations between our categorizations of decision-makers with extreme preferences exceed 0.6 across all categories, with a maximum of 0.81 for patient and a minimum of 0.64 for risk-neutral decision-makers.

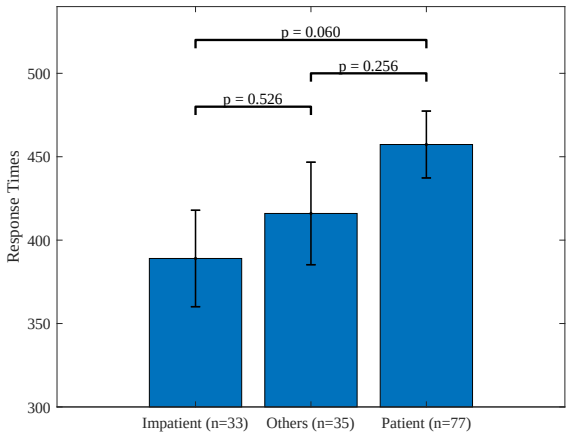


Figure 15: Response times (ordinal approach).

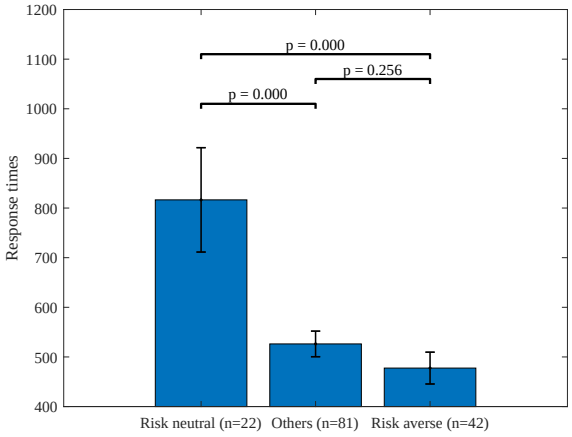


Figure 16: Response times (ordinal approach).

Notes: the bar plots report the sum of the response times in all questions, and the participants are grouped by their preferences. 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.

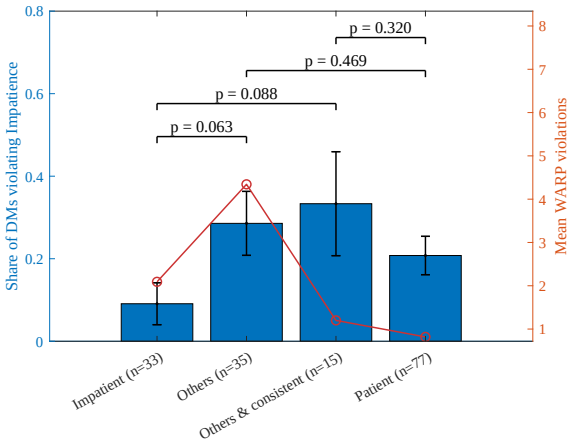


Figure 17: Share of participants violating IMP (ordinal approach).

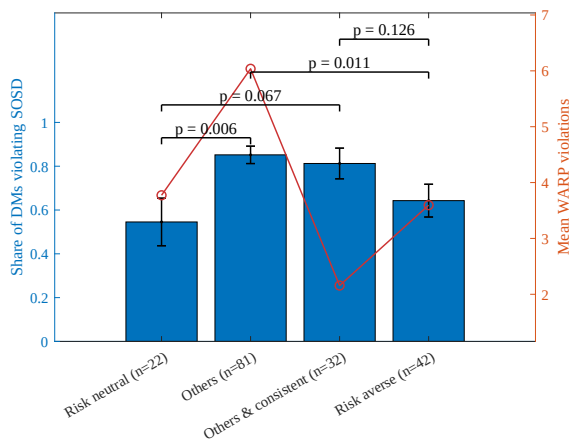


Figure 18: Share of participants violating SOSD (ordinal approach).

Notes: the bar plots report the share of participants who violate IMP and SOSD within each group defined by the preferences. The group "consistent" is defined as those participants with WARP violations lower or equal to 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.

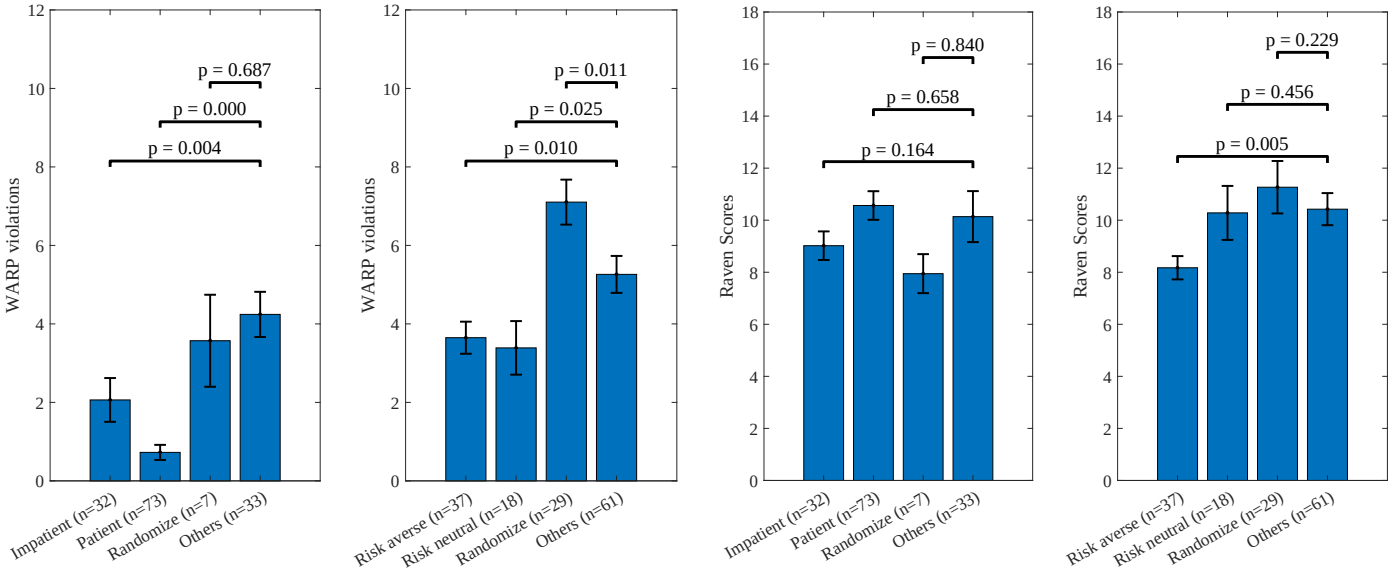


Figure 19: WARP violations (ordinal approach).

Figure 20: Raven scores (ordinal approach).

Notes: the two plots on the left report the average number of WARP violations and the two on the right 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.

3.5 Robustness check IV: Random Parameter Model

Another approach that circumvents some drawbacks of the cardinal content of the utility functions is based on the Random Parameter Model [RPM] (Apesteguia & Ballester, 2018). Here, stochastic variation occurs along the parameter dimension rather than the utility dimension. We refer to this paper for the formal definitions. For completeness, we only note that not all alternatives are optimal for some interval of the parameter space, i.e. the **50** lottery in Risk and **K** payment plan in Time (this followed a conscious decision as these alternatives are simpler than the remaining ones, and so they were chosen as preferred by a significant fraction of decision-makers in our pilot). To guarantee that each alternative has a positive probability of being chosen in every menu, we augment the model with a tremble error, see Appendix B.C and C in Apesteguia & Ballester (2018) for details about the estimation of RPMs with more than two alternatives and with a tremble error.

The table reports the correlation between the estimates of the preference parameters, and the subsequent figures replicate our analysis in the paper.

	Exp	Hyp	$\beta - \delta$	Exp (RPM)
Exp	1.000	0.998	0.993	0.936
Hyp	0.998	1.000	0.989	0.938
$\beta - \delta$	0.993	0.989	1.000	0.936
Exp (RPM)	0.936	0.938	0.936	1.000

	CRRA	CARA	CRRA/ce	CARA/ce	CRRA (RPM)
CRRA	1.000	0.977	0.892	0.896	0.752
CARA	0.977	1.000	0.948	0.950	0.781
CRRA/ce	0.892	0.950	1.000	0.995	0.857
CARA/ce	0.896	0.950	0.995	1.000	0.853
CRRA (RPM)	0.752	0.781	0.857	0.853	1.000

Table 6: Correlations between preference parameter estimates

Notes: we report the linear correlations between the individual preference parameter estimates under different utility specifications (as in the paper), and we add the estimates with the Random Parameter Model.

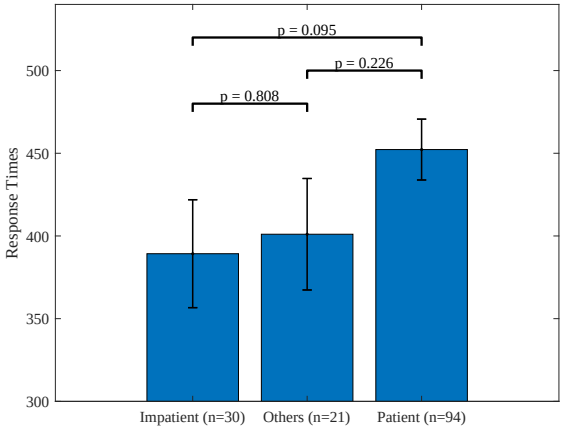


Figure 21: Response times (RPM).

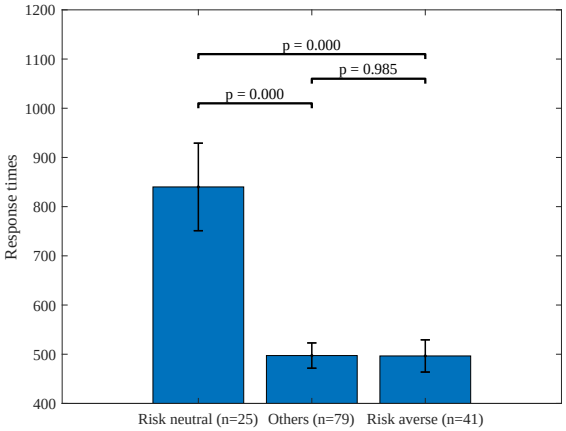


Figure 22: Response times (RPM).

Notes: the bar plots report the sum of the response times in all questions, and the participants are grouped by their preferences. 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.

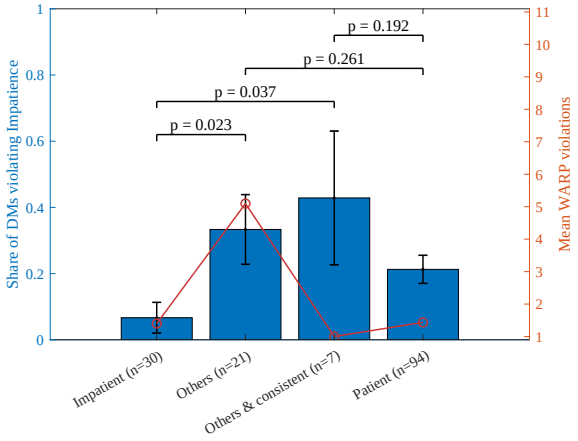


Figure 23: Share of participants violating IMP (RPM).

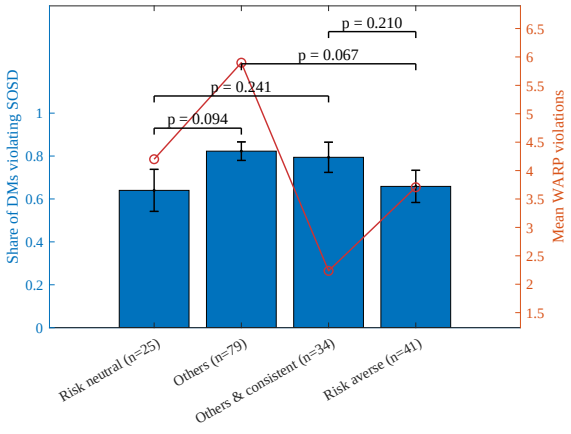


Figure 24: Share of participants violating SOSD (RPM).

Notes: the bar plots report the share of participants who violate IMP and SOSD within each group defined by the preferences. The group "consistent" is defined as those participants with WARP violations lower or equal to 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.

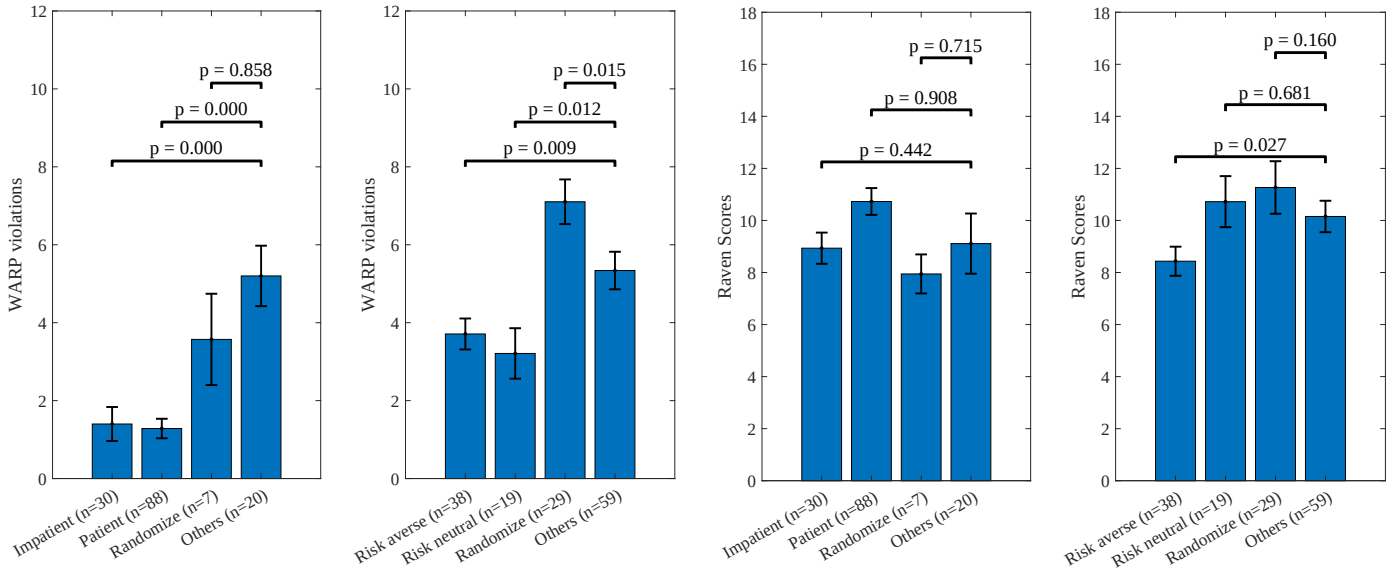


Figure 25: WARP violations (RPM).

Figure 26: Raven scores (RPM).

Notes: the two plots on the left report the average number of WARP violations and the two on the right 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.

3.6 Robustness check V: the definition of extreme preferences

In the paper, we adopted a specific definition of extreme preferences, i.e. the first and last deciles of the parameter space. Non-parametric methodologies and the (unstructured) cardinal approach provide robustness checks for more stringent definitions of "extremeness". Here, we provide a robustness check for less stringent definitions, in particular, the 2nd/8th and 3rd/7th deciles of the parameter space. We replicate the main figures from the paper.

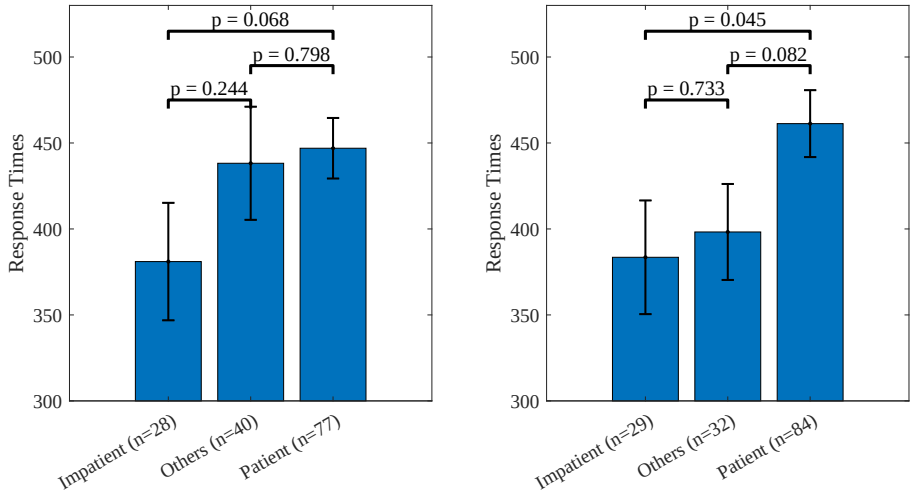


Figure 27: Response times in Time.

Notes: the bar plots report the sum of the response times in all questions, and the participants are grouped by their preferences. The left panel regards the 2nd/8th deciles, the right panel the 3rd/7th deciles. 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.

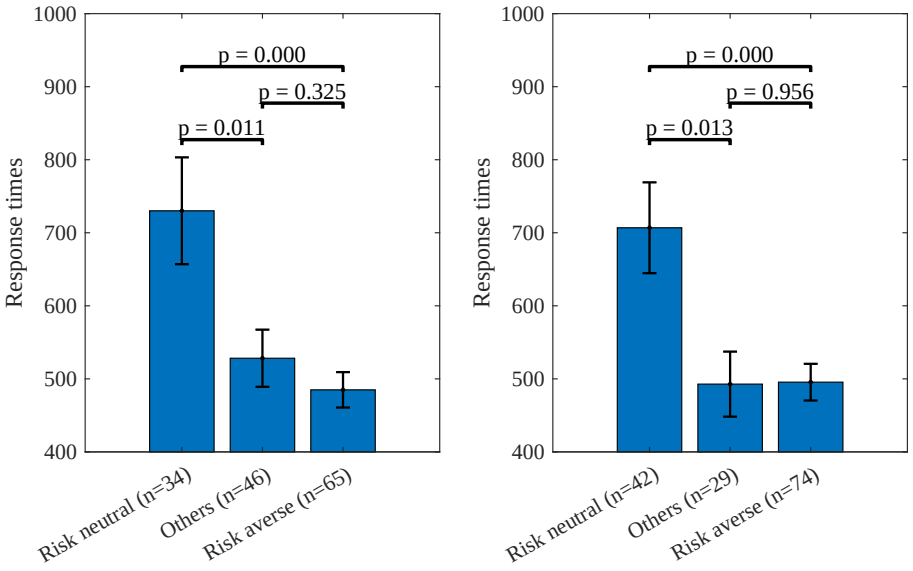


Figure 28: Response times in Risk.

Notes: the bar plots report the sum of the response times in all questions, and the participants are grouped by their preferences. The left panel regards the 2nd/8th deciles, the right panel the 3rd/7th deciles. 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.

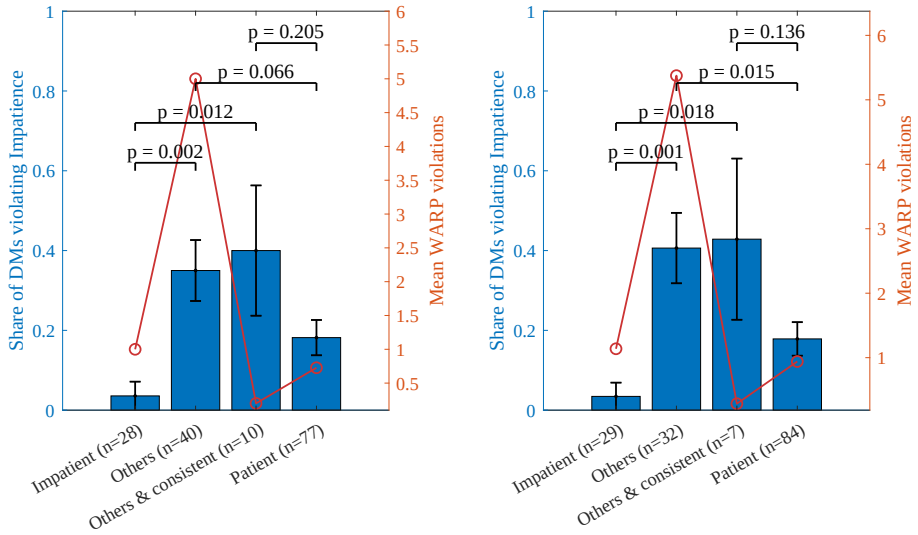


Figure 29: Share of participants violating IMP in Time.

Notes: the bar plots report the share of participants who violate IMP within each group defined by the preferences. The left panel regards the 2nd/8th deciles, the right panel the 3rd/7th deciles. The group "consistent" is defined as those participants with WARP violations lower than or equal to 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.

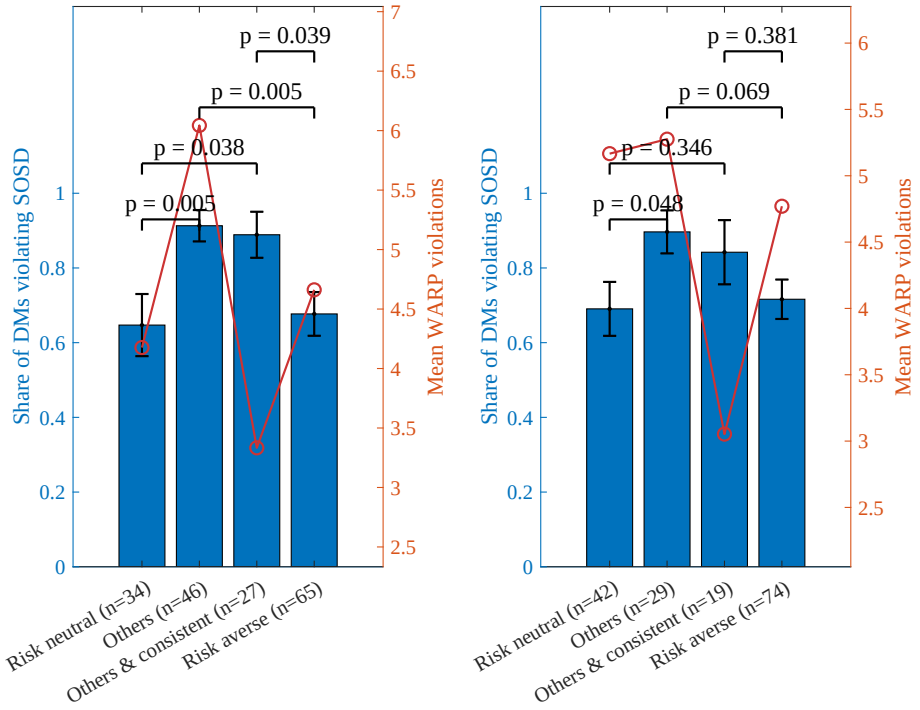


Figure 30: Share of participants violating SOSD in Risk.

Notes: the bar plots report the share of participants who violate SOSD within each group defined by the preferences. The left panel regards the 2nd/8th deciles, the right panel the 3rd/7th deciles. The group "consistent" is defined as those participants with WARP violations lower than or equal to 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.

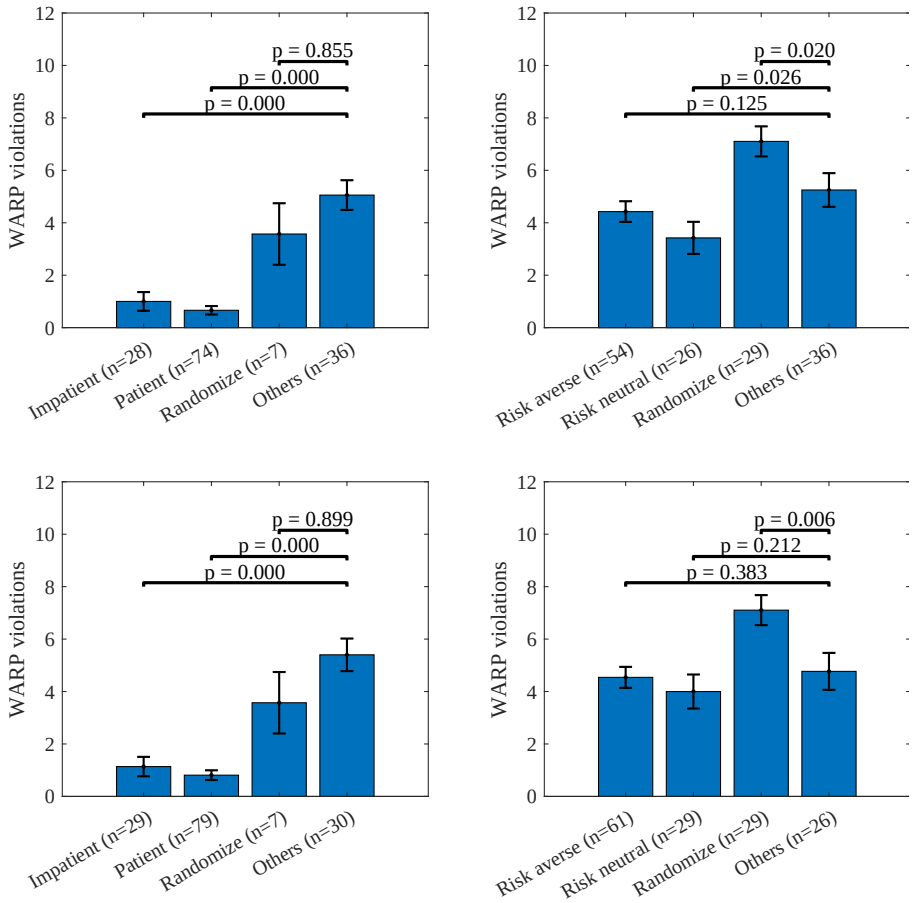


Figure 31: WARP violations and Raven scores in Time (alternatives extreme preferences).

Notes: the plots on the left report the average number of WARP violations and the two on the right report the average Raven scores for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). The top panels regard the 2nd/8th deciles, the bottom panels the 3rd/7th deciles. 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.

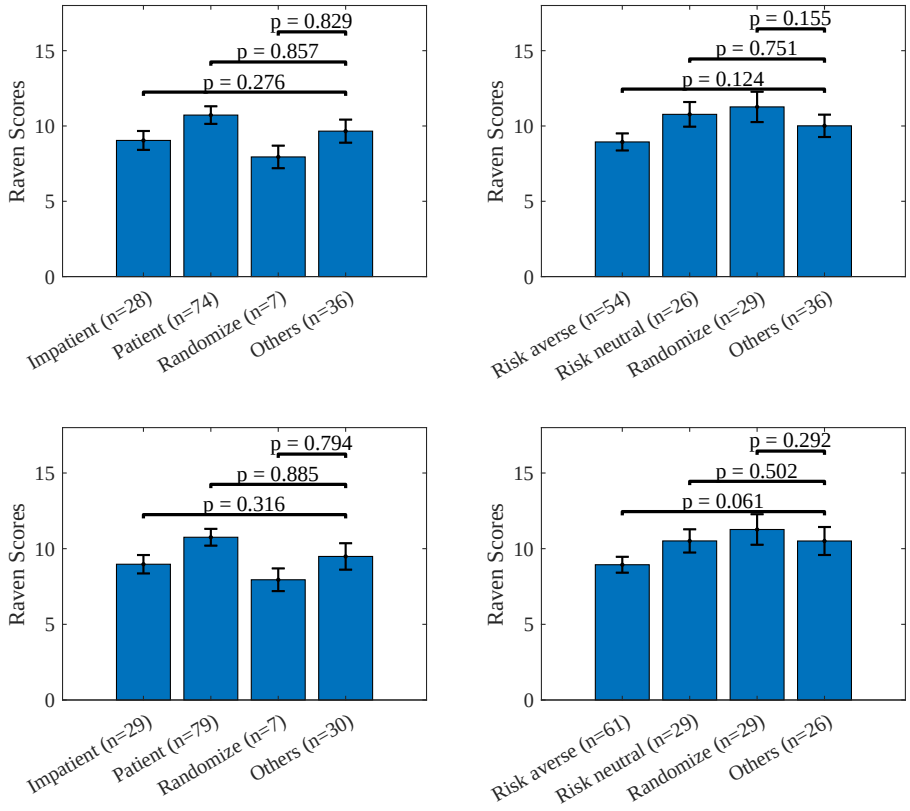


Figure 32: WARP violations and Raven scores in Risk (alternative extreme preferences).

Notes: the plots on the left report the average number of WARP violations and the two on the right report the average Raven scores for each group ("Impatient", "Patient", "Risk averse", "Risk neutral", "Randomize"), and the remaining participants ("Others"). The top panels regard the 2nd/8th deciles, the bottom panels the 3rd/7th deciles. 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.

3.7 Robustness check VI: WARP violations, Raven scores, and response times under revealed preference techniques

In previous versions of the paper, we employed an optimal weighting algorithm (Caliari, 2023) to reveal the participants' preferences. For completeness, we define this methodology and replicate the results. The dataset is divided into five parts: binary (B), ternary (T), quaternary (Q), asymmetric dominance (AD), and big (BIG) menus. This partition is clarified in Tables 1 and 2. The definition of revealed preference is then as follows

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

where $OW_{xy} = \sum_{j\in\Gamma} w_j C^j_{xy}$ and $\Gamma = \{B, T, Q, AD, BIG\}$.

The term "optimal weighting" refers to the maximization problem that yields a weight for each part of the dataset. We consider a researcher who has to identify the most preferred alternative and/or the entire preference for each participant for policy purposes, and we use the reported preferences as the benchmark. If the alternative/preference is not uniquely identified, by the principle of insufficient reason, she picks from the set of maximal alternatives with a uniform distribution. This gives us an expected proportion of identified participants for whom either (1) the most preferred alternative or (2) the entire welfare preference is uniquely identified. We call these measures **EI** ("expected identification") and **EWRI** ("expected welfare relation identification"). Formal definitions are available in Caliari (2023).

From here, the optimal weights are calculated by optimizing the sum of the expected identification of the most preferred alternative (EI) and the entire welfare preference (EWRI) as follows:

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

where for each participant i:

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

The resulting optimal weights are reported in Caliari (2023). Here, we limit ourselves to robustness checks. In the following subsections, we report results the compare our structural estimation with the all non-parametric methodologies, as well as the reported preferences and behaviour in the questionnaire: (i) structural estimation; (ii) questionnaire; (iii) Minimum Swaps (Apesteguia & Ballester, 2015); (iv) sequential method (Horan & Sprumont, 2016); (v) reported preferences; (vi) optimal weighting method (Caliari, 2023).

3.7.1 WARP violations

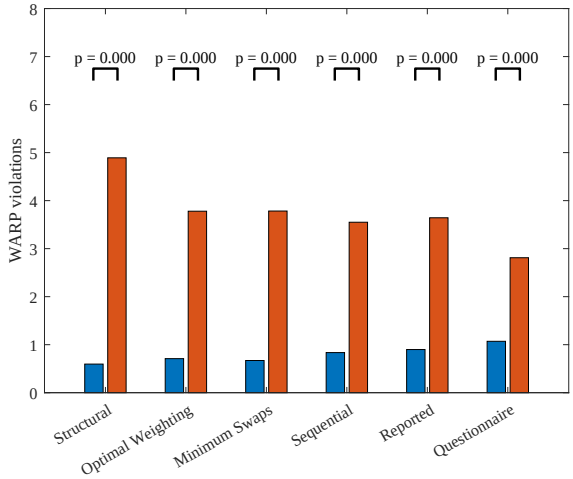


Figure 33: Violations of WARP in Time.

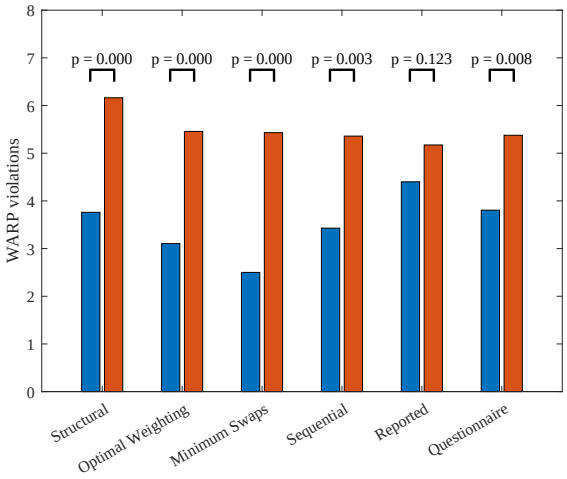


Figure 34: Violations of WARP in Risk.

Notes: the bar plots report the average number of WARP violations for participants with extreme preferences (blue bars) vs the other participants (orange bars). 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.

3.7.2 Raven scores

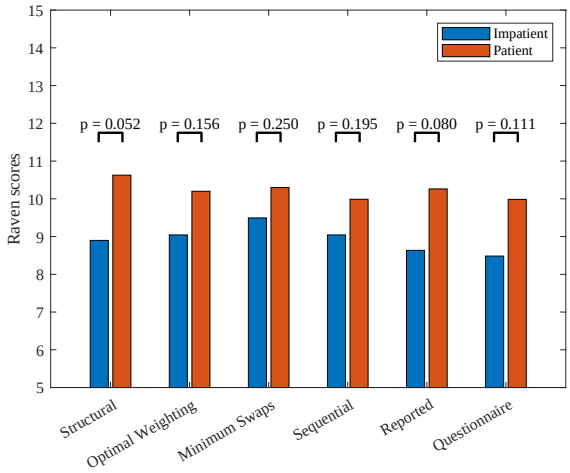


Figure 35: Raven scores in Time.

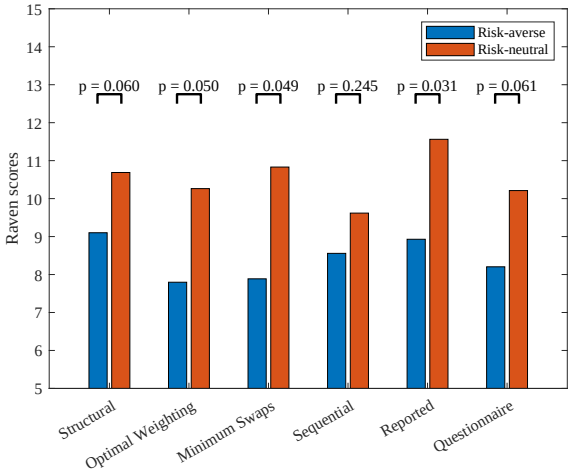


Figure 36: Raven scores in Risk.

Notes: the bar plots report the average Raven score for participants with extreme preferences (blue bars) vs the other participants (orange bars). 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.

3.7.3 Response Times

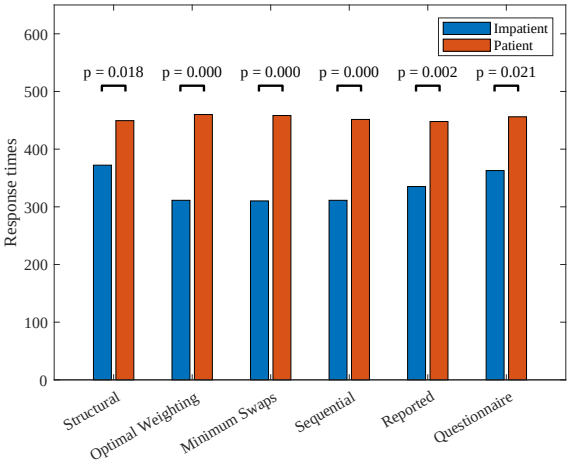


Figure 37: Response times in Time.

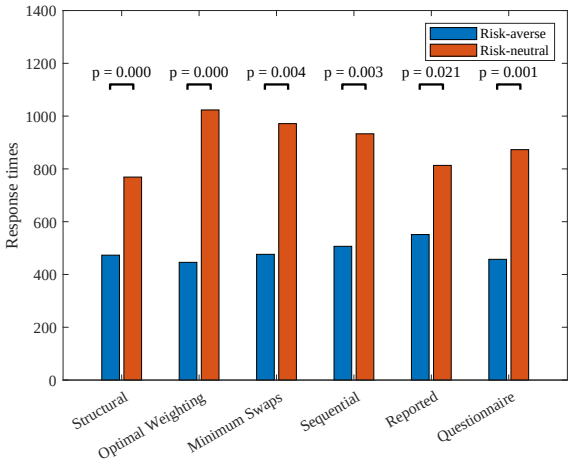


Figure 38: Response times in Risk.

Notes: the bar plots report the average response times for participants with extreme preferences (blue bars) vs the other participants (orange bars). 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.

3.8 Stationarity and Independence

Expected and discounted utility theories are founded on two **ConAx** that impact the functional form of the utility function without limiting the decision-makers' preferences in absolute terms. These axioms can be written in the form: "If you choose **a** from the menu **A**, then you should choose **b** from the menu **B**". In Time the axiom is Stationarity (STA) which is generally defined as follows: for all $a \in \mathbb{R}$ and $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ if you prefer $(a, x_2, \ldots, x_n)$ to $(a, y_2, \ldots, y_n)$ then you also prefer $(x_2, \ldots, x_n, a)$ to $(y_2, \ldots, y_n, a)$. STA induces the common exponential discounting

functional form to the discounted utility. In our framework, STA implies that if a is chosen from $\{OS, D, K, I\}$ then $\mathbf{Im}(a)$ is chosen from $\{\mathbf{Im}(OS), \mathbf{Im}(D), \mathbf{Im}(K), \mathbf{Im}(I)$ (questions 11 and 16 in Table 3).

In Risk, the axiom is Independence (IND): for all lotteries l, l', if you prefer l to l' then you should also prefer $\alpha l + (1-\alpha)l''$ to $\alpha l' + (1-\alpha)l''$ for every lottery l'' and $\alpha \in [0,1]$. It is well-known that IND induces the linearity in probabilities that characterizes expected utility theory. Unlike STA, our experiment is not designed to provide a perfect test of IND. However, an imperfect test for IND can be performed by comparing choices between $\{S, 50, R\}$ and $\{\mathbf{FOSD}(S), \mathbf{FOSD}(50), \mathbf{FOSD}(R)$ (questions 10 and 18 in Table 4).

Before looking at correlations, some important observations are due. First, since we are testing STA and IND along the impatience and FOSD dimensions, we focus on decision-makers who do not violate these properties. Secondly, we expect the following results: (i) violations of WARP may impact violations of STA and IND, hence we should observe a positive correlation between these violations; (ii) since randomization (Cerreia-Vioglio et al., 2019) is directly founded on convexity, these decision-makers should violate IND more often than the remaining ones; (iii) due to randomization we expect violations of IND to be more prevalent than those of STA.

We confirm all predictions. We find positive correlations between violations of WARP and STA in Time (r = 0.32, p-value < 0.01) and between violations of WARP and IND in Risk (r = 0.26, p-value < 0.01). We also find that only 28% of randomizers satisfy IND while 53% of the remaining participants do so (Fisher exact test, p-value < 0.05). Finally, in our population, 75% of participants satisfy STA while only 48% satisfy IND.

Finally, in Figure 39, we show the correlation between cognitive abilities and violations of STA and IND. The results show no significant correlations.

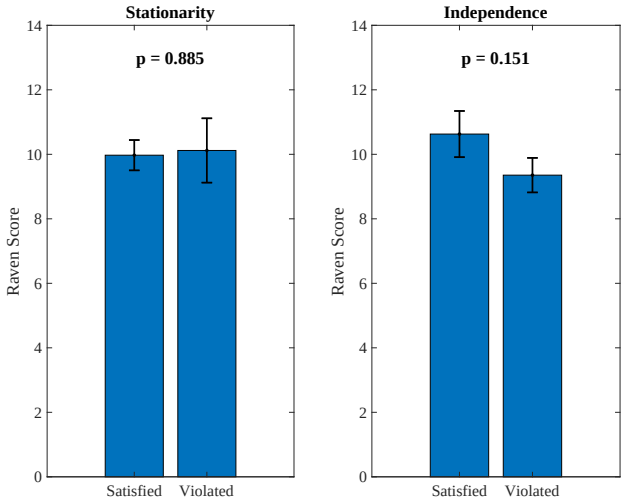


Figure 39: Violations of stationarity and independence and Raven's scores.

Notes: the bar plots report the average Raven scores for participants who did and did not violate the preferences axioms STA and IND. Above the bars, we include the p-values from Welch's t-tests.

3.9 Robustness check VII: measurement of consistency

A potential source of variation in our results is the measurement of consistency. In Figures 40 and 41, we replicate our analysis on WARP violations using the Swaps index and the Houtman-Maks index. First of all, the weak and ambiguous correlations between consistency and cognitive abilities are confirmed with these indexes: for the Swaps index (in Time, r = -0.07, p-value = 0.39; in Risk, r = 0.08, p-value = 0.31) and for the Houtman-Maks index (in Time, r = -0.06, p-value = 0.48; in Risk, r = 0.04, p-value = 0.68). All results are expected because the Swaps and

Houtman-Maks indexes are highly correlated with WARP violations: 0.97 and 0.96 in Time, and 0.96 and 0.93 in Risk, respectively. Below, in Figures 40 and 41, we confirm our results on the relation between heuristics, randomization, and consistency.

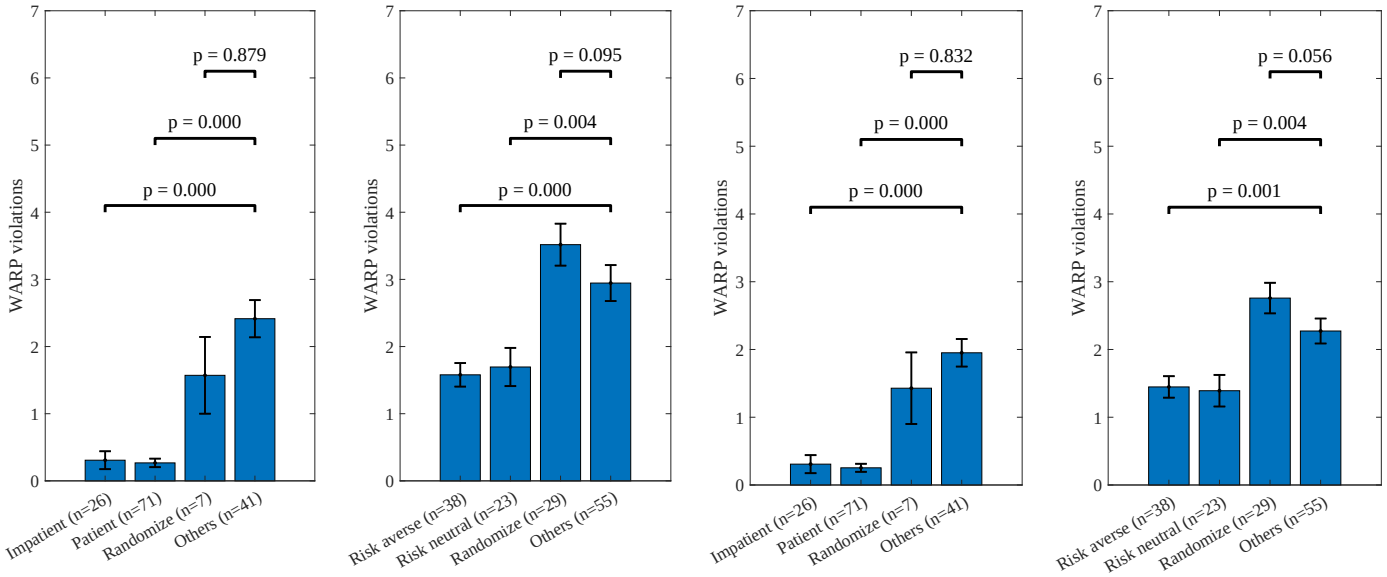


Figure 40: Swaps Index (heterogeneous analysis).

Figure 41: Houtman-Maks Index (heterogeneous analysis).

Notes: the two plots on the left report the average Swaps index and the two on the right report the average Houtman-Maks index 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.

Both the Houtman-Maks index and the Swaps index are distances from utility maximization, and in fact, they are highly correlated with the number of violations of WARP. We now focus on two conditions that are weaker than WARP but jointly equivalent. We explore whether WARP is too strong to be a necessary condition for high decision-making ability, while these weaker properties may be. The two properties are: No Binary Cycle and Always Chosen. We refer to Manzini & Mariotti (2007) for the formal definitions. Again, we confirm the ambiguous correlations with cognitive abilities: for No Binary Cycle (in Time, r = 0.08, p-value = 0.32; in Risk, r = -0.07, p-value = 0.44) and for Always Chosen (in Time, r = -0.15, p-value = 0.08; in Risk, r = 0.10, p-value = 0.24). Note that these indices are less correlated with WARP violations: 0.47 and 0.82 in Time, and 0.34 and 0.73 in Risk, respectively. Below, in Figures **??** and **??**, we again confirm our results.

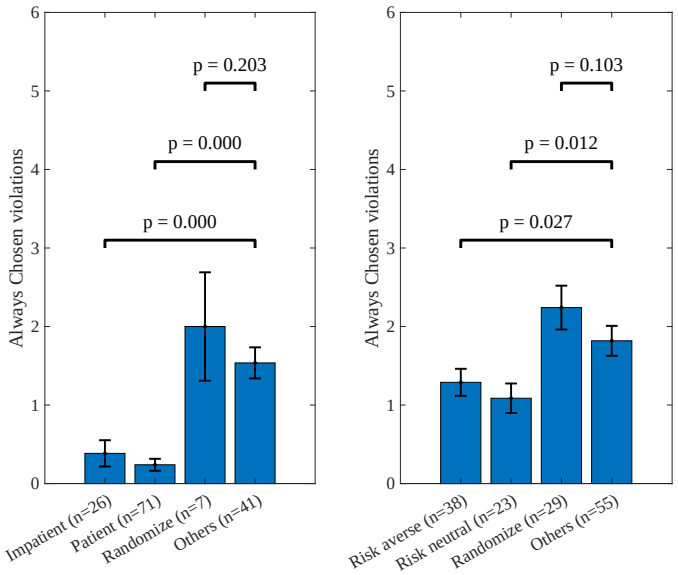


Figure 42: Always Chosen (Manzini & Mariotti, 2007) (heterogeneous analysis).

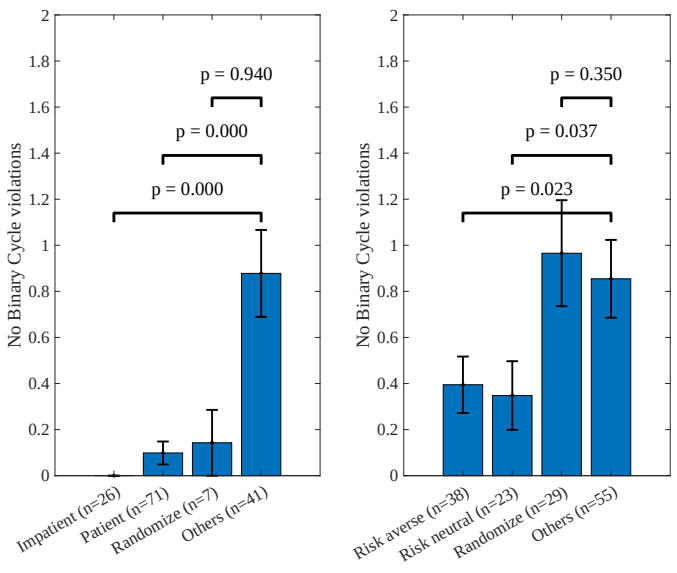


Figure 43: No Binary Cycle (Manzini & Mariotti, 2007) (heterogeneous analysis).

Notes: the two plots on the left report the average number of violations of Always Chosen and the two on the right report the average number of violations of No Binary Cycle 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.

3.10 Robustness check VIII: measurement of cognitive abilities

This section discusses the measurement of cognitive abilities. The subset of the Raven test used comprises 10 questions of varying difficulty. We construct a measure as a weighted average of the indicator function that indicates whether the answer is correct. We denote this measure as r^w. The definition is as follows:

$$r^w = \sum_{q \in \text{Raven}} \mathbf{1}_q \cdot w_q$$

where $\mathbf{1}_q$ is equal one if the answer to the question q is correct and zero otherwise. The weight w_q is the inverse of the percentage of participants who correctly answered question q. In the following figures, we replicate the results of a non-weighted version of the Raven scores and the Cognitive Reflection Test (Frederick, 2005). The two measures are correlated at 0.26 ($p < 0.01$). The results are similar in Risk, while in Time, the CRT yields slightly different results, but with no significant differences. Note that the difference between risk-averse and randomizers remains significant (p-value < 0.05) using the non-weighted Raven, while it is not significant (p-value > 0.1) using the CRT.

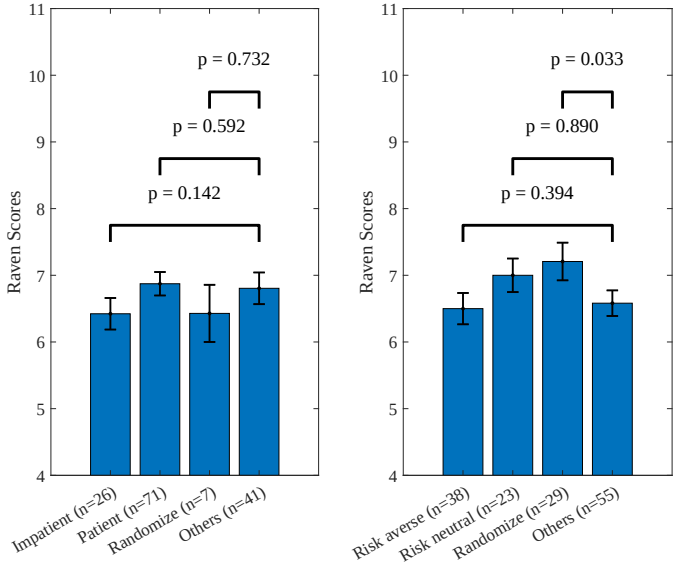


Figure 44: Non-weighted Raven

Notes: the plots report the average non-weighted Raven scores for each group ("Impatient", "Patient", "Risk averse", "Risk neutral"), randomization ("Randomize"), and the remaining participants ("Others"). The categorization is based on the structural estimation in the paper. 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.

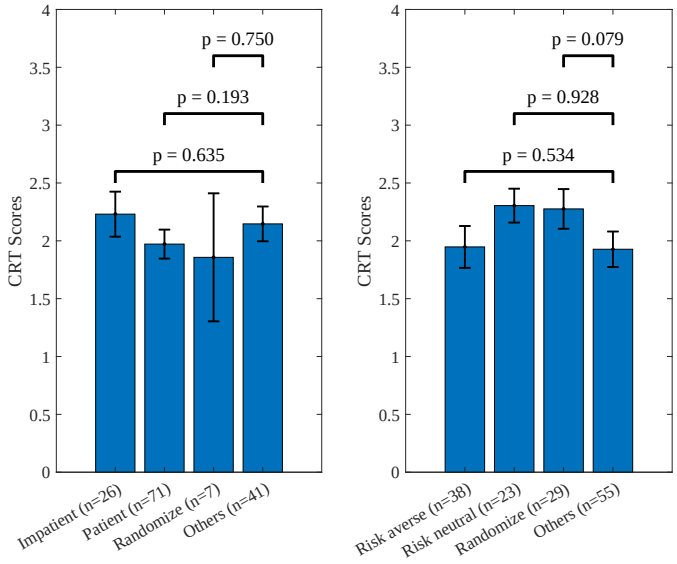


Figure 45: Cognitive Reflection Test

Notes: the plots report the average Cognitive Reflection Test scores for each group ("Impatient", "Patient", "Risk averse", "Risk neutral"), randomization ("Randomize"), and the remaining participants ("Others"). The categorization is based on the structural estimation in the paper. 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.

3.11 Robustness check IX: test of RUM without heuristics and randomization

In the paper, we documented two failures of ARUM using response times and violations of Impatience and SOSD. We have also identified heuristics and randomization as drivers of these failures. To provide more evidence in favour of this interpretation, we repeat our tests excluding participants with extreme preferences and those who self-reported randomization behaviour. We show that among the remaining participants, ARUM is no longer rejected. To do so, we regress response times and violations of Impatience and SOSD on our estimates of the discount factor and risk parameter, with a squared component, because our hypotheses involve U-shaped forms. We first include and then exclude these participants, comparing the results. In table 7, Columns 1 and 5 replicate the results in the main text, showing that patient and risk-neutral participants are slower. Columns 2 and 6 report the same regressions but excluding participants with extreme preferences and randomizers. The monotonic relation is no longer significant. Columns 4 and 8 test our hypothesis that response times should have a U-shaped form, but we find no evidence for it. Similarly, in table 8, Columns 1 and 5 replicate the results in the main text, showing a clear U-shaped relation between violations of impatience, SOSD, and the preference parameters. Again, Columns 2 and 6 find no evidence of such a relation in the restricted subsample. Finally, Columns 4 and 8 test our hypothesis of a monotonic relation, again, though the direction is correct, we find no significant evidence.

Note that since the change in the significance of the results could be due to the smaller sample, once we exclude part of the participants, we run 10000 bootstrap regressions with the smaller sample size (41 in Time and 55 in Risk) but using the entire sample and report the mean coefficients, which obviously are close to our original estimate, and the p-values for their difference from zero. Columns 3 and 7 in Tables 7 and 8 confirm that our results are not simply due to the smaller sample size of the subsamples.

VARIABLES	(1) Resp. Times	(2) Resp. Times	(3) Resp. Times	(4) Resp. Times	(5) Resp. Times	(6) Resp. Times	(7) Resp. Times	(8) Resp. Times
Discount factor	732.9**	790.2	714.68	-129,340				
	(356.2)	(1,997)	p = 0.28	(147,375)				
Discount factor^2				68,200				
				(77,147)				
Risk parameter					-107.2***	-59.72	-107.83**	-346.2
					(31.43)	(59.21)	p = 0.018	(323.4)
Risk parameter^2								106.8
								(116.6)
Heuristics & Deliberate Rand.	✓	×	✓	×	✓	×	✓	×
Observations	145	41	41	41	145	55	55	55

Table 7: Chronometric function

Notes: the dependent variable is the sum of response times, and the independent variables are preference parameters. The regression models are estimated in Stata (Robust standard errors). *** <0.01, ** <0.05, *<0.1.

VARIABLES	(1) Impatience	(2) Impatience	(3) Impatience	(4) Impatience	(5) SOSD	(6) SOSD	(7) SOSD	(8) SOSD
Discount factor	221.5***	306.3	225.69*	1.547				
	(68.64)	(228.3)	p = 0.09	(3.738)				
Discount factor^2	-115.8***	-159.7	-118.03*					
	(36.15)	(120.7)	p = 0.093					
Risk parameter					0.665***	0.547	0.667**	-0.0415
					(0.181)	(0.528)	p = 0.0312	(0.0882)
Risk parameter^2					-0.277***	-0.219	-0.279**	
					(0.0733)	(0.199)	p = 0.0308	
Heuristics & Deliberate Rand.	✓	×	✓	×	✓	×	✓	×
Observations	145	41	41	41	145	55	55	55

Table 8: Monotonicity of violations of Impatience and SOSD

Notes: the dependent variable is the number of violations of Impatience, and the independent variables are preference parameters. The regression models are estimated in Stata (Robust standard errors). *** <0.01, ** <0.05, *<0.1.

4 Questionnaire

To gain a deeper understanding of students' perceptions of the experiment, we conducted a non-incentivized questionnaire. In the first part, summarized in Table 9, participants were asked to agree or disagree with a set of statements. We find that participants show both a good understanding of the experimental design and the instructions. They also report evidence of a learning effect expressed both in terms of preference learning and quicker response times. The reward has been considered a significant contribution to their overall daily budget.

	QUESTIONS	MEAN	SD
1	**"I got a good understanding of overall experiment (incentives, how to choose, nature of the alternatives)".**	1.552	0.600
2	**"The instructions and explanations provided have been enough to understand experiment's duties".**	1.366	0.622
3	**"The clarity of my preferences on the alternatives improved during the course of the experiment".**	1.993	0.961
4	**"The time required for choosing has reduced during the course of the experiment".**	1.931	1.045
5	**"The money I earn through participating in this experiment is a substantial contribution to my daily budget"**	2.462	1.167

Table 9: Questionnaire 1

In the second part of the questionnaire (Table 10), participants were asked to reply with a "Yes" or "No" to a series of questions about the structure of the experiment and their behaviour. Overall, they acknowledge that some questions were more difficult than others, and they indicate the high number of alternatives as the main source of complexity. Strangely, most participants had the impression that some questions were repeated, even though this was not the case. They overall confirm that they have read all alternatives before choosing. Finally, they confirm the presence of reference point effects.

Questions 6 and 7 focus on the reasoning of participants. About half of the participants report risk neutrality via the use of the "highest expected value" criterion, while about two-thirds report

patience via the use of the "highest summation" criterion.

	QUESTIONS	YES	NO
1	**"Did you find that some questions were more difficult than others?"**	123	22
2	**"If yes, what features did play a main role in making these questions more difficult?" [YES - High number of alternatives] - [NO - Complexity of some alternatives]**	92	53
3	**"Do you think some questions were asked multiple times?"**	133	12
4	**"Did you always read all the alternatives carefully before choosing?"**	118	27
5	**"Did you base some decisions on previous choices?"**	123	22
6	**"In case of lotteries, in general, did you calculate the expected value and choose according to the "highest expected value" criterion?"**	75	70
7	**In case of delayed payment plans, in general, did you calculate the summation of the plan and choose according to the "highest summation" criterion?**	97	48

Table 10: Questionnaire 2

4.1 Samples of open answers on heuristics and randomization

Patient

- "Highest summation every time."
- "Added the tokens up, and chose the highest."
- "always went for the choice that over the long run gave the greatest income"
- "I'm not in desperate financial circumstances, therefore it was objectively better for me to go for the highest payoff over the whole time period."
- "I calculated which would have the highest return, regardless of how long I would have to wait for the money. "
- "Whichever option has the highest number of tokens overall"
- "As the interest rate is very small and inflation as well, I just calculated the sum"

Most Impatient

- "I strongly preferred payment plans that paid most or all of the total amount today, even if the total amount was smaller than other payment plans that involved long delays."
- "highest payment in time 0."
- "It's better to get money now since money tends to lose its value within time. In UK after BREXIT there is a high risk of increasing inflation, therefore money I get in 12 months may cost nothing."
- "The number of tokens involved were often so small that I preferred an immediate payment over a delayed one, even if a delayed payment would have paid me a little more."

- "Earlier payment"
- "How fast the payment will be transferred (inflation was also a criterion)"

Randomization (Time)

- "I did usually calculate the summation of the plan, but sometimes chose a plan which would be paid more quickly."
- "I considered the total payment as a factor but if a high fraction of the largest total payment option was available very quickly then I was tempted to switch."
- "Oftentimes used highest summation criterion, but also sometimes went for the instant gratification option due to the relatively low amount of money at stake."
- "Mainly (particularly with the first 15 plans)calculated highest summation and chose it. Later, if summations were within 40 tokens, I sometimes went with alternatives w/ earliest large payment."
- "In general, I summed the payment plans to find highest payout. Occasionally chose the lower payout option if the highest payout for a particular time was 12 months."

Most Risk Averse

- "least risky."
- "I chose the options that were the safest but still offered a decent amount of tokens such as 100% probability for 50 tokens since this guaranteed tokens."
- "most likely to win something. less risk."
- "Highest probability to win."
- "whatever seemed like I was mostly likely to get a decent payout (eg £5) with minimal risk of getting nothing"

Risk Neutral

- "highest expected value."
- "Calculate the expected value of the lotteries and choose the highest expected value."
- "Calculate expected value, note down. If a question only contains seen lotteries, look up previous values. I may have chosen some 300|15 when I thought they were 300|20."
- "Disciplined myself into being risk-neutral and chose highest expected value lotteries even though it was hard to choose a less secure option"
- "Sometimes expected value was too hard to calculate without a calculator, so I tried to guess what the highest value would be. "

Randomization (Risk)

- "I like to take risk to get the best payment, but sometimes I will take a middle one."
- "Highest likelihood of a reasonable gain. Once or twice I went big with high-risk, high-reward options."
- "in some questions, I make decisions based on expected value while in some other questions, I prefer the certain gain."
- "I switched between avoiding getting 0 and taking the risk - 300 tokens in first part became too enticing to keep choosing the safe options"
- "I mixed my answers to include both the high risk lotteries and the low risk one's that guaranteed tokens"
- "I was relatively risk averse. If I was guaranteed 50 Tokens I would take that over a higher expected outcome with greater risk, although sometimes I chose a risky option."
- "Wanted to make sure I could get the largest possible prize even if I won the lower value. In some questions, I decided to take a gamble"

5 Instructions

General Instructions

This is an experiment in the economics of decision-making and in particular risk and time attitudes of individuals.

Notice there is no RIGHT or WRONG answer for any of these questions. We are interested in studying your preferences.

The tasks are extremely simple and if you make good decisions you may earn a **considerable amount of money** that will be **paid to you by direct debit** at the end of the experiment and at particular times in the future.

The currency in this experiment is called **tokens.**

The experiment should last at maximum 1:30 hour even though more probably around 1 hour. It is divided in four parts:

1. **Part 1:** 25 Questions;
2. **Part 2:** 25 Questions;
3. **Questionnaire**
4. **Test**

At the end of experiment, the computer will randomly choose on question for each part of the experiment and pay your choice. In case of Lottery the computer will also play out the Lottery according to the respective probabilities.

Your total earning from the experiment will consist of the sum of three components:

1. One Choice from Part 1;
2. One Choice from Part 2;
3. A participation fee of £ 5.

Notice:

- There is no time constraint for any question;
- Before each part of the experiment you will complete a three questions trial in order to make you fully understand which kind of questions you are going to answer. The alternatives in the trial questions do not have sense and will not be considered in the calculation of your reward.
- After you click the "OK" button, there might be a short delay before the next question appears, due to the software. Please be patient.
- At the end of **PART 1** the screen **"BREAK"** will appear. Please remain sit and wait. During the break you are not allowed to speak with anyone. You will receive the instructions for **PART 2** and the experiment will restart.

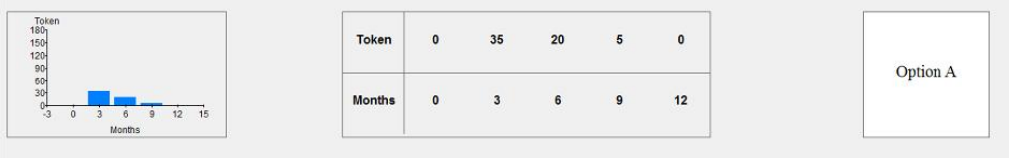
Specific Instructions

Instructions are visualized on the screen before each Part of the Experiment. This is the paper version. You can keep it and revise it in case you have doubts about the experiment.

The exchange rate for this part of the experiment is:

20 Token = £ 1

In this part, you will choose among Delayed Payment Plans. The following screen shows a delayed payment plan:

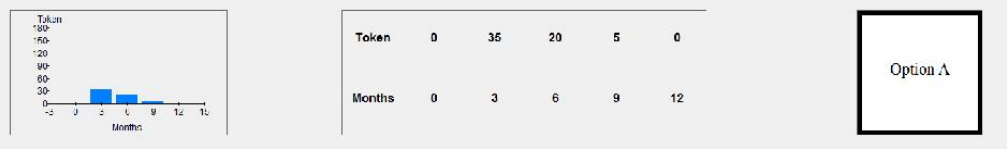


A delayed payment plan is described in two ways:

- A histogram, on the left, allows to visualize the number of tokens paid at different months;
- A table in which Token and Months are written.

This delayed payment plan allows you to win 35 Tokens in 3 months, 20 Tokens in 6 months and 5 Tokens in 9 months.

On the right you can see a Button to select. If you click on “Option A” you select this Delayed Payment Plan and a black frame will appear.



You can change your choice even after you have selected one. When you are sure of your choice you can click the Button “OK” in the Right-Bottom part of the screen (see below). Once you clicked “OK” your decision is recorded and you cannot change it.

Notice:

- Questions may have a different number of alternatives to choose from;
- Delayed Payment Plans may different both in Token and in Time of Payments.

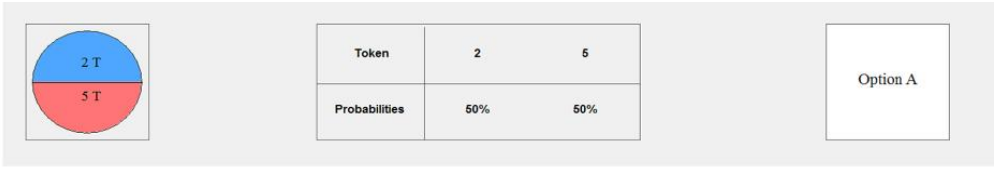
Specific Instructions

Instructions are visualized on the screen before each Part of the Experiment. This is the paper version. You can keep it and revise it in case you have doubts about the experiment.

The exchange rate for this part of the experiment is:

10 Token = £ 1

In this part, you will choose among Lotteries. The following screen shows a lottery:

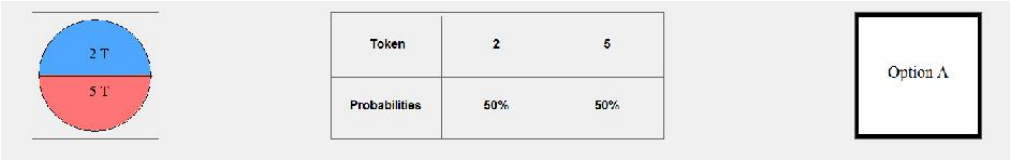


A lottery is described in two ways:

- A pie, on the left, allows to visualize the probabilities to win;
- A table in which Token and Probabilities are written.

This lottery allows you to win 2 Token with 50% probability and 5 Token with 50% probability.

On the right you can see a Button to select. If you click on “Option A” you select this Lottery and a black frame will appear.



You can change your choice even after you have selected one. When you are sure of your choice you can click the Button “OK” in the Right-Bottom part of the screen (see below). Once you clicked “OK” your decision is recorded and you cannot change it.

Notice:

- Questions may have a different number of alternatives to choose from;
- Lotteries may different both in Token and in Probabilities.

6 Screens

Figure 46: Screenshot of a question in Risk

Figure 47: Screenshot of a question in Time

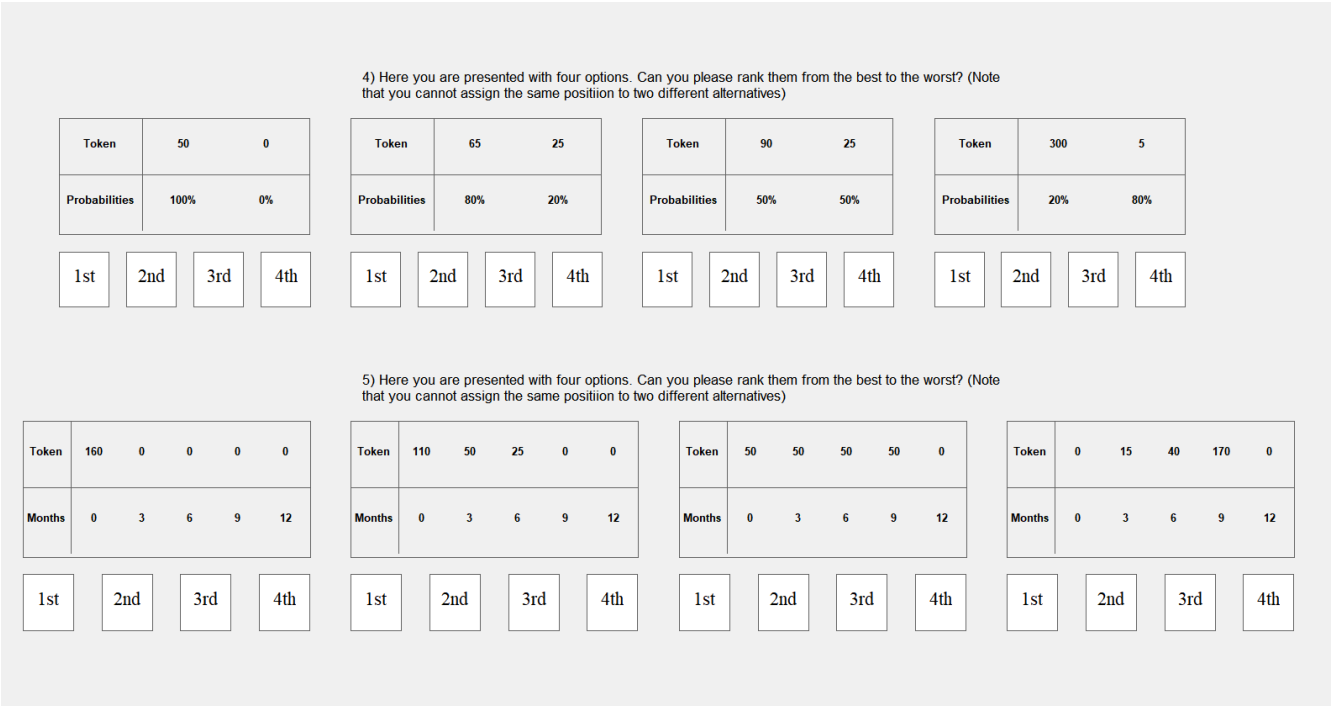


Figure 48: Screenshot of the elicitation of reported preferences

References

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

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

Apesteguia, J., & Ballester, M. A. (2018). Monotone stochastic choice models: The case of risk and time preferences. *Journal of Political Economy*, *126*(1), 74–106.

Apesteguia, J., Ballester, M. A., & Lu, J. (2017). Single-crossing random utility models. *Econometrica*, *85*(2), 661–674.

Badescu, B., & Weiss, W. (1987). Reflection of transitive and intransitive preferences: a test of prospect theory. *Organizational Behavior and Human Decision Processes*, (39), 184–202.

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

Caliari, D. (2023). Behavioural welfare analysis and revealed preference: Theory and experimental evidence. Tech. rep., WZB Discussion Paper.

Carlsson, F., Morkbak, M. R., & Olsen, S. B. (2012). The first time is the hardest: a test of ordering effects in choice experiments. *Journal of Choice Modelling*, *2*(5), 19–37.

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.

Cerreia-Vioglio, S., Dillenberger, D., Ortoleva, P., & Riella, G. (2019). Deliberately stochastic. *American Economic Review*, *7*(109), 2425–2445.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: between-subject and within-subject design. *Journal of Economic Behavior and Organization*, (81), 1–8.

Day, B., Bateman, I., Carson, R., Dupont, D., Louviere, J., Morimoto, S., Scarpa, R., & Wang, P. (1987). Ordering effects and choice set awareness in repeat-response stated preference studies. *Journal of Environmental Economics and Management*, *1*(63), 73–91.

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

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

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

Ladenburg, J., & Olsen, S. B. (2008). Gender-specific starting point bias in choice experiments: evidence from an empirical study. *Journal of Environmental Economics and Management*, *3*(56), 275–285.

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

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

McCausland, W. J., Davis-Stober, C., Marley, A., Park, S., & Brown, N. (2019). Testing the random utility hypothesis directly. *The Economic Journal*, (130), 183–207.

Plott, C. R. (1993). Rational individual behavior in markets and social choice processes. *Working paper - California Institute of Technology*.

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