The risk elicitation puzzle in a representative sample: A potential resolution*

Jorgo T.G. Goossens† Marike Knoef‡ Bart Kuijpers§
Rogier Potter van Loon¶ Eduard H.M. Ponds‖ Arno Riedl**
Siert Vos††

September 28, 2022

*We thank Rob van den Goorbergh and seminar participants at Netspar, APG, APG Asset Management, and bpfBOUW Pension Fund for useful comments.

†Radboud University, Institute for Management Research, Heyendaalseweg 141, 6525AJ, Nijmegen. Corresponding author: jorgo.goossens@ru.nl

‡Leiden University, department of Economics

§APG Asset Managent, Research and Analytics

¶Aegon

‖Tilburg University, department of Economics

**Maastricht University, Department of Microeconomics and Public Economics

††PGB Pensioendiensten

The risk elicitation puzzle in a representative sample: A potential resolution

Abstract

Numerous methods to elicit and classify people's risk attitudes have evolved and evidence suggests that risk preferences may vary considerably when measured with different methods. Based on a within-subject design using three widespread risk preference elicitation methods, we find that the different methods indeed give rise to considerably varying estimates of risk preferences. Thus, we confirm the existence of the "risk elicitation puzzle" in our sample. By conducting simulation exercises, consistent with earlier findings, we show that part of the observed heterogeneity of risk preferences across tasks is due to task-specific measurement error induced by the mechanics of the task. By additionally using representative unique data from 1601 pension fund members, we show that the observed heterogeneity in risk preference (in)stability across elicitation methods is similar to the heterogeneity arising from at least 50% of the sample making at random choices in the elicitation methods. Therefore, we provide a potential resolution for the "risk elicitation puzzle". We are the first to show that risk preference stability is more prevalent for individuals that are white collar workers, younger, high income earners, and risk tolerant. We do not find statistical and economic differences between design attributes of elicitation tasks, i.e., the format of presenting the tasks in a serious game or a classic questionnaire.

Keywords: risk preference, elicitation methods, representative survey, (in)stability, pensions

JEL Codes: C90, D10, D91

Experimental evidence suggests that individuals' attitudes towards risk may vary considerably when measured with different elicitation methods. A finding recently referred to as the "risk elicitation puzzle" (Pedroni et al., 2017). What is particularly challenging about the risk elicitation puzzle is not the heterogeneity in risk preferences across different methods per se, but rather the understanding of what drives the observed variation in risk attitudes. To get a better understanding of the observed variability in revealed risk preferences across methods, we use a representative sample and simulations in this paper.

This paper addresses the following two research question. First, how stable, or consistent, are risk preferences across elicitation methods in a representative sample? Second, what drives the observed (in)stability of revealed risk preferences? Third, how does (in)stability correlate with socio-demographic variables?

We use a within-subject design compromising three widely used risk preference elicitation methods: (i) a single choice list (Eckel and Grossman, 2002; Eckel and Grossman, 2008), (ii) a choice sequence list (Barsky et al., 1997), and (iii) Convex Time Budgets (Andreoni and Sprenger, 2012). Typically, former studies that assess risk preference instability have used student samples (Holzmeister and Stefan, 2021; Crosetto and Filippin, 2016) or specific ages between 20 and 36 (Pedroni et al., 2017; Frey et al., 2017). We use a non-student sample of pension fund participants that is representative for a large pension fund in The Netherlands. Our sample is diverse in terms of characteristics, as it includes the young and elderly as well as low and high incomes. The hypothetical questions in our online experiment are directly context relevant and concern large stakes as our subjects make decisions regarding their pension payments.

We run a simulation exercise, similar to Crosetto and Filippin (2016), to study the behavior of risk preferences in the three elicitation methods and, subsequently, to study the

stability of preferences. Risk attitudes are a latent construct that can only be indirectly and imperfectly measured, and the degree of measurement error is possibly influenced by the characteristics of the elicitation methods. The simulations allow us to study the behavior of risk preferences in the elicitation methods as we reveal how the mere mechanics of the tasks influence the risk preference estimates, imposing that no behavioral artifacts, like framing effects, enter the picture. We find that the different methods do introduce systematic task-specific measurement errors.

While previous studies typically assess the variability of risk preferences based on correlations, we use individual-level measures based on implied CRRA parameter intervals and based on the order of CRRA parameters in the overall distribution. Both measures confirm that stability of preferences is low; absolute stability of risk preferences is only about 10% (for the CRRA-parameter interval measure). Conducting simulation exercises, our main result is that the observed heterogeneity in revealed risk preferences is similar to the heterogeneity arising from 50% random choices per elicitation method.

Our paper makes the following three findings. First, we confirm the existence of the "risk elicitation puzzle" (Pedroni et al., 2017) in a representative sample making context relevant choices. Elicited risk aversion parameters vary in terms of level and order. Second, if we introduce 50% random choices per elicitation method for a group of virtual subjects, then we are able to match the actual observed stability in our experimental sample of pension fund participants. This provides a potential resolution for the risk elicitation puzzle: in a representative sample, subjects might simply be confused, do not understand the task well, or do not know their preferences in some methods. Third, we find that risk preference stability is more prevalent for individuals that are white collar workers, younger, high income earners, and risk tolerant. This could yield important policy implications.

Our paper contributes to the studies on preference (in)stability and preference (in)consistency (Holzmeister and Stefan, 2021; Pedroni et al., 2017; Frey et al., 2017; Crosetto and Filippin, 2016). The primary goal of our study is not per se to add to the pile of evidence of seemingly inconsistent behavior in risk elicitation methods, but rather to contribute to the understanding of the observed across-method variation in risk preferences. Specifically, it is not known yet (to the extent of our knowledge) how the mechanics of the choice sequence and CTB tasks affect risk preference estimates. And, we are the first to study the instability of risk preferences in a large diverse representative sample through the argument of random choice behavior.

1. Survey design

Our survey contains three elicitation methods. We use three frequently used elicitation methods for measuring risk preferences in the following order: a single choice list (Eckel and Grossman, 2002; Eckel and Grossman, 2008), (ii) a choice sequence list (Barsky et al., 1997), and (iii) Convex Time Budgets (Andreoni and Sprenger, 2012). After each elicitation method follows a question about how certain the individual is regarding her answers for the elicitation method. The complete survey can be found in the Online Appendix. We field our survey at one of the largest pension funds in The Netherlands; the pension fund takes care of the pensions for the construction industry.

Our survey was not incentived based directly on the answers given by the participants.[1] It was stated at the beginning of the survey that the participant's choices make her pension better and more personal, so participation in the survey was consequential. Pension funds

[1]Some researchers argue that answer-based incentives in economic experiments lead to more truthful reveal of preferences, however Cohen et al. (2020) and Hackethal et al. (2022) find little evidence for systematic differences between incentivized and unincentivized risk preference experiments.

in The Netherlands by law will be required to elicit risk preferences from their population and use it in the formation of their asset allocation. The monetary amounts shown in the survey are tailored towards the average income of the pension fund's population.

A. Risk elicitation methods

Choice sequence list - In the choice sequence list, subjects are asked to choose between two pensions: a risky and a non-risky pension. A pension is defined as a lottery. The variation is obtained through manipulations of the outcomes of each lottery, while keeping the probability of the two outcomes fixed at 50% (i.e., similar as a coin toss for heads and tails). Per lottery, one outcome contains a high payout and is defined as the situation 'better than expected', while the other outcome contains a low payout and is defined as 'worse than expected'. Subjects are asked to choose one pension per question for a total of five sequential questions. The method is based on the original approach of Barsky et al., 1997. The pensions that an individual can choose from depend on the individual's previous choices, so that risk aversion is narrowed down to a specific interval. Table 7 in the Appendix shows the values used in the experiment.

Single choice list - In the single choice list, subjects are asked to choose a pension out of an ordered set of pensions. A pension is defined as a lottery. We use a version based on the question proposed by Eckel and Grossman (2002) and Eckel and Grossman (2008). Subjects choose the preferred pension among a set of 4 lotteries characterised by a linearly increasing expected value as well as greater standard deviation (except for Pension 4). More risk averse subjects choose low risk, low return pensions; risk-neutral subjects choose Pension 3; risk-seeking subjects choose Lottery 4.[2]

[2]In our analysis, we take the average value of the interval as a proxy for the risk aversion value. On the

The variation in questions is obtained through manipulations of the outcomes of each lottery, while keeping the probability of the two outcomes fixed at 50% (i.e., similar as a coin toss for heads and tails). Per lottery, one outcome contains a high payout and is defined as the situation 'better than expected', while the other outcome contains a low payout and is defined as 'worse than expected'. Subjects are asked to choose one pension. Table 8 in the Appendix presents the values used in the experiment.[3]

Convex Time Budgets - An important advantage of the CTB is that it allows to measure risk and time preferences simultaneously. We simultaneously measure risk aversion and patience. Our approach is a shorter version of the original approach of Andreoni and Sprenger (2012), as we exclude the measurement of present bias.

The method asks individuals to allocate an initial budget $m = €10,000$ between payments, available at two points in time: an early payment at time t and a delayed payment at time $t + k$. The early payment is always one year from the experimental date (to avoid interference with present bias). The late payment is delayed by either five years $k = 5$ or ten years $k = 10$. Subjects receive an interest rate, or return, r on delayed payments, which varies between 0% to 21.06% interest on an annual basis. The allocations must be made such that their budget constraint is satisfied, i.e., the early payment and the present value of the delayed payment must equal the initial budget m. Early and late payments are certain.

Individuals make 6 consecutive CTB decisions between early and delayed payments. Our method consists of two different decision sets, and within each set we have three different interest rate scenarios. The first choice set uses $k = 10$, and the three decisions within this set differ in the accrued return. The second choice set uses $k = 5$, , and the three

bounds, we assume follows of $\gamma = 5.5$ and $\gamma = -2$.

[3]The range and cutoff points for the CRRA parameter values are based on insights from an earlier risk preference study at a Dutch pension insurer.

decisions within this set differ in the accrued return accordingly as well. Table 9 in the Appendix presents an overview of our experimental CTB design. Differences between the delayed payment dates $t + k$ (i.e., back-end delay) elicit long-term patience. Sensitivity to variation in the interest rates, or return, identifies curvature of the utility function.

To estimate risk and time preferences, we identify the experimental allocated payments as solutions to standard intertemporal optimization problems. These solutions are supposed to be functions of our parameters of interest (discounting and risk aversion) and experimentally varied parameters (interest rates and delay lengths). Given assumptions on the functional form of utility and the nature of discounting, this setup provide a natural context to jointly estimate individual.

We assume that the agent has a standard CRRA utility function with curvature parameter γ and that the agent is a exponential discounter with discount factor δ. We estimate risk and time preferences together, i.e., the CRRA risk aversion parameter γ and the long-term discount factor δ. Our preference estimates are based on OLS regressions.[4] In line with the former literature (Andreoni and Sprenger, 2012; Potters et al., 2016) and consistent with our two previous elicitation methods, we assume a background income close to zero. That is, we make the assumption that participants do not integrate any other income sources with their CTB decisions. See Goossens and Knoef (2022) for more details on the estimation.

[4]The preference estimates are robust to using TOBIT specifications.

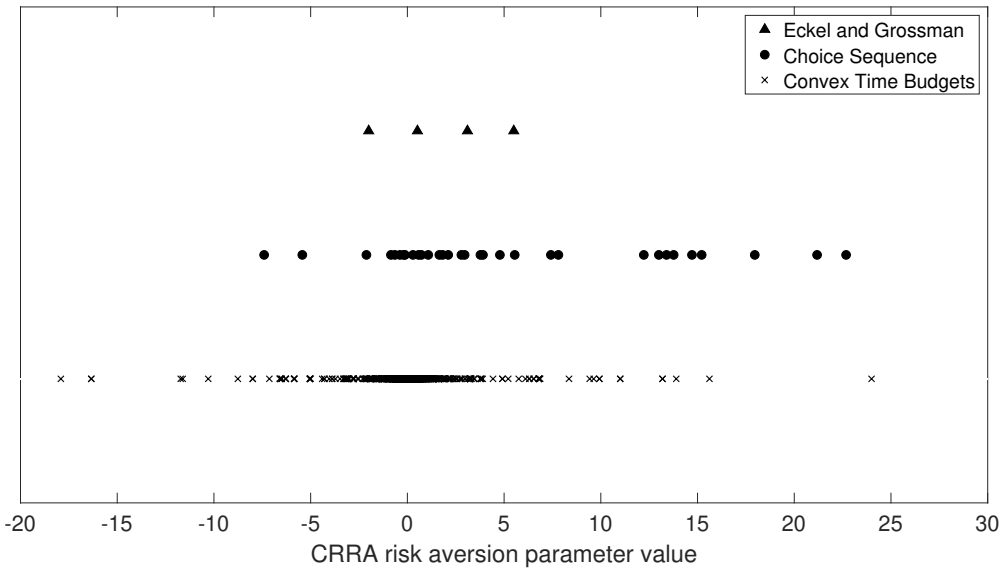


Figure 1: **Mapping of choices into the implied CRRA risk aversion parameter value by task.** The figure assumes a CRRA power function $u(x) = \frac{x^{1-\gamma}}{1-\gamma}$. $\gamma = 0$ means risk neutral, $\gamma > 0$ implies risk-averse preferences, and $\gamma < 0$ implies risk-seeking preferences.

B. Theoretical possibilities

Figure 1 shows all theoretically possible outcomes for each task in the space of the CRRA coefficient γ. We assume throughout the paper that the utility function is of the CRRA form

$$U(x) = \frac{x^{1-\gamma}}{1-\gamma}, \tag{1}$$

where $\gamma = 0$ implies risk-neutral behavior, $\gamma > 0$ implies risk-averse behavior, and $\gamma < 0$ implies risk-seeking behavior. It is immediately clear from the figure that EG is a coarse measure with only four outcome possibilities. CS has 32 possible distinct values, and CTB has 4096 possible distinct values. CTB, and to a lesser extent CS, are thus continuous measures. Notably, CTB is well suited for distinguishing between risk seeking, risk neutral, and risk averse behavior, since many outcome possibilities cluster around zero.

Table 1: **Comparison experimental sample and actual pension fund.** Mean values with standard deviations between brackets. Annual before-tax income for employed participants actively accruing a pension (i.e., excluding retirees).

	Experimental sample $N = 1601$	Pension fund
Male	0.95	0.92
	(0.23)	
Age (years)	59	55
	(12)	
Income (Euros)	49937	45300
	(24558)	

C. Sample

Table 1 shows that our experimental sample of $N = 1601$ pension fund participants has similar characteristics as the total pension fund population. The distribution of males is very skewed, as most construction workers are male. The average age and income in our experimental sample is a bit higher than at the pension fund level, but lie well within one standard deviation. The median time to complete the survey is 5 minutes.

2. Results

Figure 2 summarises part of the main evidence for the confirmation of the risk elicitation puzzle in our sample, i.e., the instability of risk preferences across methods. Distributions of the elicited risk aversion parameters differ substantially between methods. Note that in our analyses all risk aversion values for all three methods are winsorized at a 1% level (bottom and top). Table 2 presents additional summary statistics for these experimental results. The median risk aversion estimates γ are 3.11, 5.55, and 0.23 for EG, CS, and CTB, respectively. The standard deviation in CS is the highest, 8.54, compared to 1.88 and 1.70 in CS and

CTB respectively. In the CTB about 21% is risk seeking, while only about 3% and 7% is risk seeking in EG and CS respectively. The cognitive certainty at the median is 3 for all methods on a 4-point Likert scale, which indicates that participants are equally 'sure' about there answers across all methods.[5]

Not only the levels of risk aversion differ quite substantially across methods, but also the rank order of individuals. Figure 3 shows the scatterplots along with a fitted-least squares line and the Spearman's rank correlation coefficients. The correlation coefficient between EG and CS is 0.46 and statistically significant, but the correlations between EG and CTB and CS and CTB are negligible and insignificant. Overall, this shows that risk aversion levels and rank orders differ across methods, confirming the risk elicitation puzzle.

A. Cognitive certainty

Table 3 shows an interesting positive relation between risk aversion and self-stated cognitive certainty. We sort the subjects on demeaned self-reported cognitive certainty and divide the data in four groups of equal size. Group 1 is the most cognitive uncertain group, and Group 4 is the most cognitive certain group. The table shows that cognitive uncertain subjects are consistently less risk averse across methods. The mean risk aversion is statistically different from Groups 2 and 3. The mean risk aversion of Group 4 is not statistically different from Groups 2 and 3. Significance test are done using a non-parametric Wilcoxon rank-sum test. So, a policy recommendation could be to protect people that are cognitively uncertain as they might make too offensive investment choices.

[5]This could relieve any concerns regarding the fixed order of our questions, as participants do not become more sure about their answers later in the survey.

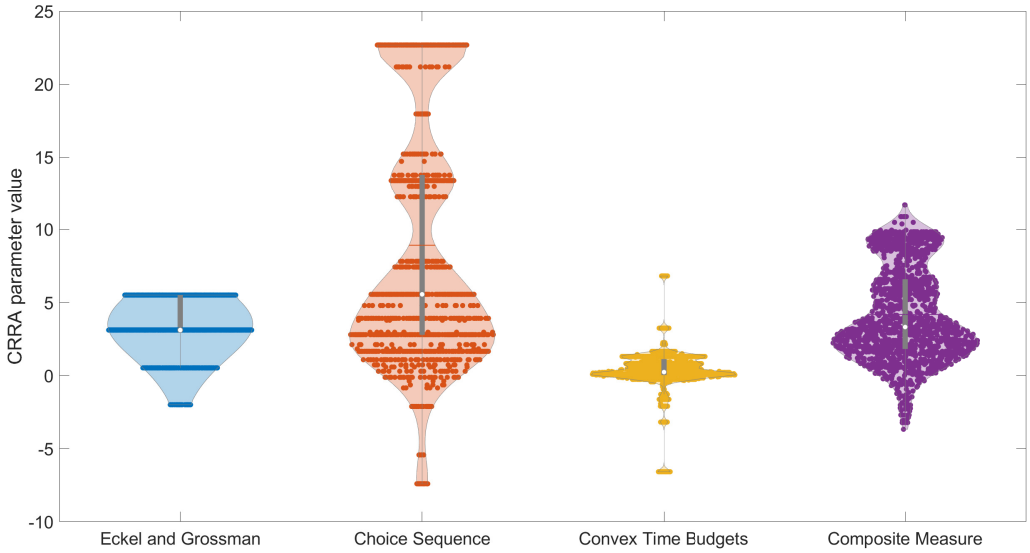


Figure 2: **Experimental results by task.** In the violin plots, the white dot in the equals the median, and the vertical grey bold bars run from the first quartile to the third quartile. On the vertical axis the CRRA risk aversion parameter value, and on the horizontal axis the elicitation method. The composite measure is an (unweighted) average of the CRRA parameter values of the three elicitation methods.

Table 2: **Summary statistics experimental results.** The composite measure is an (unweighted) average of the CRRA parameter values of the three elicitation methods. Cognitive certainty is measured on a 4-points Likert scale with 1 = 'very unsure' and 4 = 'very sure'.

	EG	CS	CTB	Composite
Type of choice	Chosen lottery	Chosen lotteries	Chosen delayed return	
Choice options	4	32	4096	
Choice set γ	[-2.0, ..., 5.5]	[-7.4, ..., 22.7]	[-308.7, ..., 376.5]	
Chosen γ				
Median	3.11	5.55	0.23	3.31
Mean	3.26	8.91	0.33	4.17
Std. dev.	1.88	8.54	1.70	3.24
Min.	-2.00	-7.41	-6.58	-3.68
Max.	5.50	22.68	6.81	11.66
% Risk seeking	2.87	6.81	20.67	5.43
Cognitive certainty				
Median	3.00	3.00	3.00	
Mean	2.97	2.87	2.91	
Std. dev.	0.67	0.70	0.68	
Observations N	1601	1601	1601	1601

Table 3: **Demeaned cognitive certainty and risk aversion.**

	Mean CRRA risk aversion			
	EG	CS	CTB	Obs
Group 1: Cognitive uncertain	3.03	7.77	0.19	321
Group 2	3.33	9.57	0.38	961
Group 4: Cognitive certain	3.32	8.08	0.33	321
Difference Group 1 and 2	0.30	1.80	0.19	
Ranksum difference test p-value	0.01	0.00	0.07	

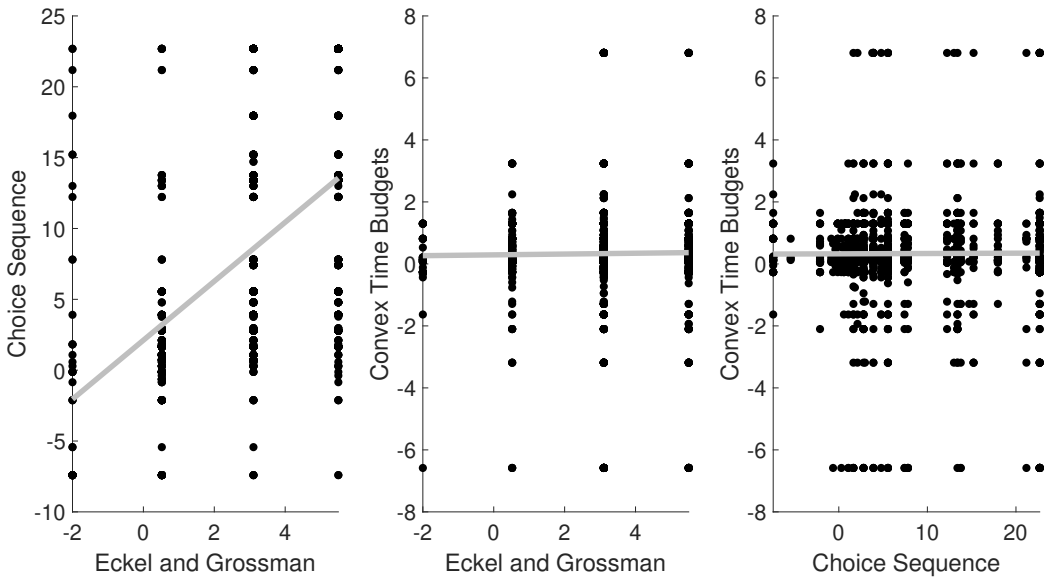


Figure 3: **Scatterplots of experimental results with fitted least-squares lines.** On both axes CRRA risk aversion parameter values for the specified elicitation methods. Spearman's rank correlation coefficients are from left to right in the figure: 0.46*, 0.01, and 0.01. * Indicates significance at 1% level.

3. Simulation

We generate 10,000 virtual subjects, each characterized by the CRRA utility function (1) with $\hat{\gamma}$ drawn from the fitted distribution in Figure 4. The (non-parametric) kernel distribution is chosen such that it generates a realistic sample given the observed values from all methods in our survey. In particular, the mean, median, and standard deviation for the simulated risk aversion equals 4.17, 2.48, and 6.29 respectively. The distribution yields most of the mass that is accounted for by risk averse subjects, with a smaller share of risk seeking subjects.

Each of these virtual subjects, with true underlying simulated risk aversion $\hat{\gamma}$, is exposed to the three elicitation tasks. Using the generated virtual risk aversion values, we let the virtual subjects 'proceed' through the elicitation methods and then retrieve the individual

coefficient of risk aversion. This procedure allows to numerically evaluate which bias in the measurement of risk preferences, if any, follows by definition from distortions generated by the mere technical features of the task.

We do three simulations, similar to Crosetto and Filippin (2016). First, we assume totally deterministic preferences. That is, each virtual subject acts exactly as their coefficient of risk aversion dictates. The observed, revealed, risk aversion γ_{obs} equals exactly their true underlying simulated risk aversion $\hat{\gamma}$. This measures the range and precision of the tasks.

Second, we assume stochastic preferences. That is, we add noise to the subject's true underlying simulated risk aversion. The observed, revealed, risk aversion γ_{obs} departs from their true underlying simulated risk aversion $\hat{\gamma}$, according to some noise. The added noise follows the fitted demeaned distribution from Figure 4. These noisy preferences may induce subjects to make a choice different than that dictated by the true $\hat{\gamma}$, e.g., subjects might not themselves know their true underlying risk aversion or due to measurement error.

Third, we assume that 10%, 50%, and 100% of the virtual subjects pick randomly a risk aversion value out of the set offered, per elicitation method, instead of following their $\hat{\gamma}$. This procedure simulates confused subjects, and measures the robustness of the elicitation methods. Thus, per elicitation method, a certain fraction of the responses is chosen at random with a uniform distribution over the theoretically possible answers.

Comparison simulations

Table 4 shows the observed risk aversion values γ_{obs} (Panel B) from the true underlying simulated risk aversion values $\hat{\gamma}$ (Panel A). Figure 5 shows the distance between γ_{obs} and $\hat{\gamma}$, based on the outcomes of Table 4: a low value (green) indicates that the observed risk

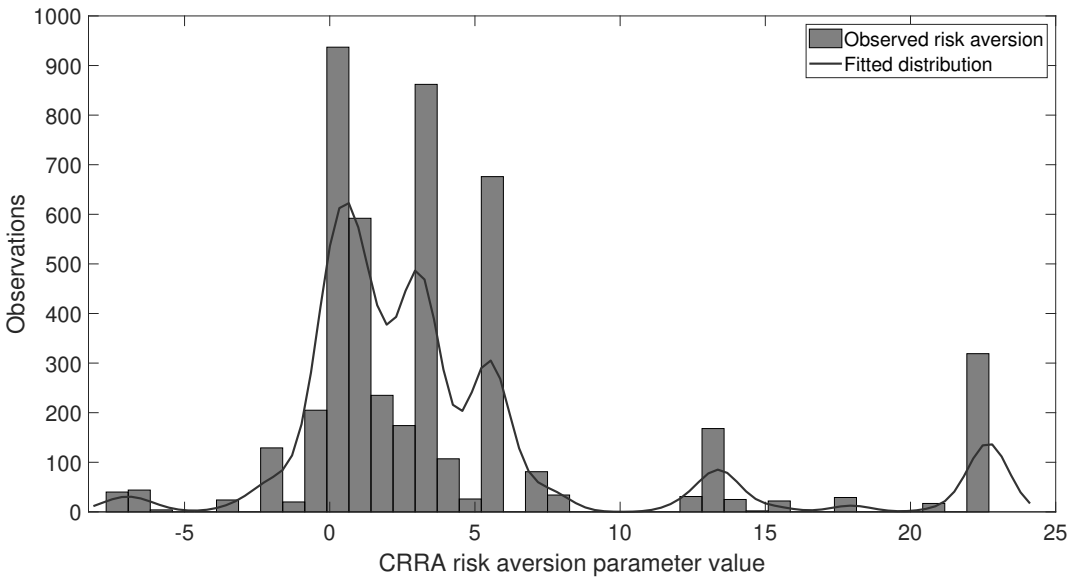


Figure 4: **Fitted distribution to observed risk aversion values.** The solid line is a fitted (non-parametric) kernel distribution with the following mean 4.17, median 2.48, and standard deviation 6.29.

aversion is close to the true underlying risk aversion.[6]

Deterministic preferences - The CTB and CS methods are superior to EG in terms of matching the level and variability of preferences in the case preferences are deterministic. The reason is that the CTB and CS methods are essentially almost continuous measures.

Stochastic preferences - Also in this case the CTB and CS methods perform relatively well in matching the level of preferences, especially the mean.

Random preferences When 50% of the subjects behave random, the CS method matches the level and variability of risk aversion well, and the EG method matches the median also well but has too low standard deviation. The CTB method pulls risk aversion towards risk neutral behavior when participants start to choose randomly and also the standard deviation increases rapidly.

[6]The figure does not display values for 100% random behavior with stochastic preferences, as it yields identical results to deterministic preferences.

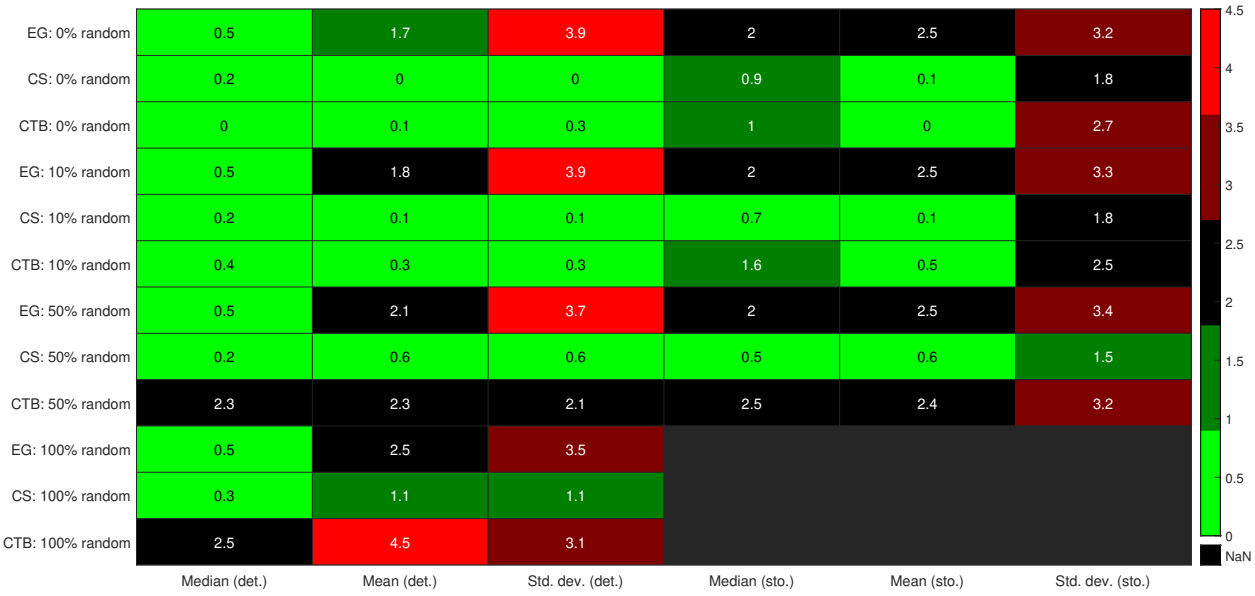


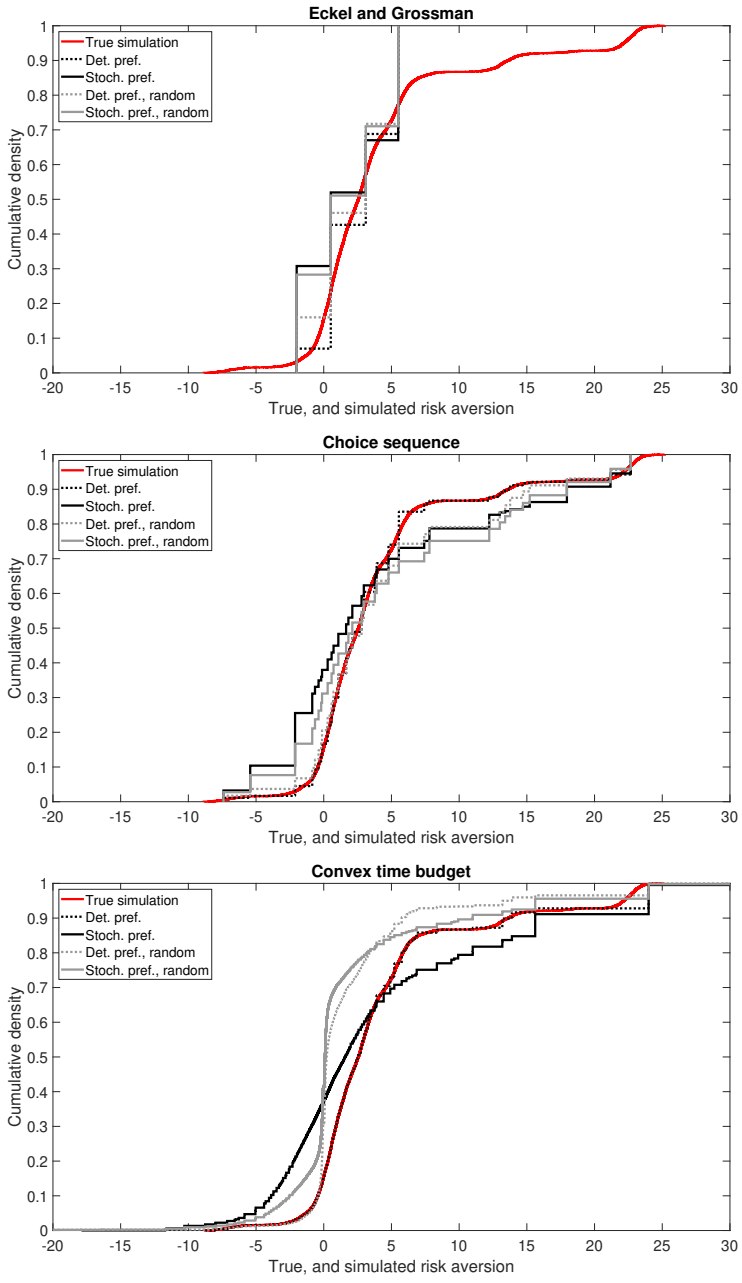
Figure 5: **Distance between true and simulated risk aversion values.** On the horizontal axis the mean, median, and standard deviation for deterministic preferences (columns 1-3) and stochastic preferences (columns 4-6). On the vertical axis the three elicitation methods grouped by the percentage of random choices.

Figure 6 summarises the above findings in the empirical cumulative density functions.

Table 4: **True, simulated, and experimentally observed coefficients of risk aversion by task.** Number of virtual subjects 10,000.

	Panel A: True underlying simulated risk aversion					
	Median	Mean	Std. dev.	Median	Mean	Std. dev.
True	2.56	4.29	6.33	2.56	4.29	6.33
	Panel B: Elicited risk aversion from virtual subjects					
	Deterministic			*Stochastic*		
	Median	Mean	Std. dev.	Median	Mean	Std. dev.
EG	3.11	2.57	2.39	0.52	1.78	3.09
CS	2.79	4.27	6.28	1.65	4.14	8.17
CTB	2.57	4.38	6.60	1.60	4.26	9.06
	Deterministic, random 10%			*Stochastic, random 10%*		
	Median	Mean	Std. dev.	Median	Mean	Std. dev.
EG	3.11	2.49	2.44	0.52	1.76	3.06
CS	2.79	4.35	6.41	1.82	4.23	8.08
CTB	2.12	3.94	6.61	0.93	3.84	8.83
	Deterministic, random 50%			*Stochastic, random 50%*		
	Median	Mean	Std. dev.	Median	Mean	Std. dev.
EG	3.11	2.19	2.64	0.52	1.77	2.96
CS	2.79	4.93	6.94	2.11	4.87	7.84
CTB	0.21	1.96	8.45	0.09	1.92	9.51
	Only random choices					
	Median	Mean	Std. dev.			
EG	3.11	1.81	2.82			
CS	2.88	5.44	7.44			
CTB	0.05	-0.20	9.42			
	Panel C: Experimentally observed risk aversion					
	Median	Mean	Std. dev.			
EG	3.11	3.26	1.88			
CS	5.55	8.91	8.54			
CTB	0.23	0.33	1.70			
Composite	3.31	4.17	3.24			

Figure 6: **Comparison between true and simulated coefficients of risk aversion by task.** Number of virtual subjects 10,000.



Comparison experimental results

Table 4, Panel C, shows the experimentally risk aversion values together with the simulated values γ_{obs} from Panel B. Figure 7 shows the distance between the experimentally observed risk aversion γ and the simulated values $\hat{\gamma}$, based on the outcomes of Table 4: a low value (green) indicates that the experimentally elicited risk aversion is close to the simulated risk aversion.

An interesting observation is that at least 50% random behavior yields simulated risk aversion values that come close to the experimentally elicited risk aversion values. Randomness is required to let the mean and median of CS differ, and to increase the variability of CS. Randomness is also required to pull CTB in the direction of the experimentally observed risk neutral behavior.

Figure 8 confirms these findings by means of the empirical CDFs. 50% random behavior (grey lines) comes close to the experimentally elicited risk aversion values (red lines) especially for CTB and CS.

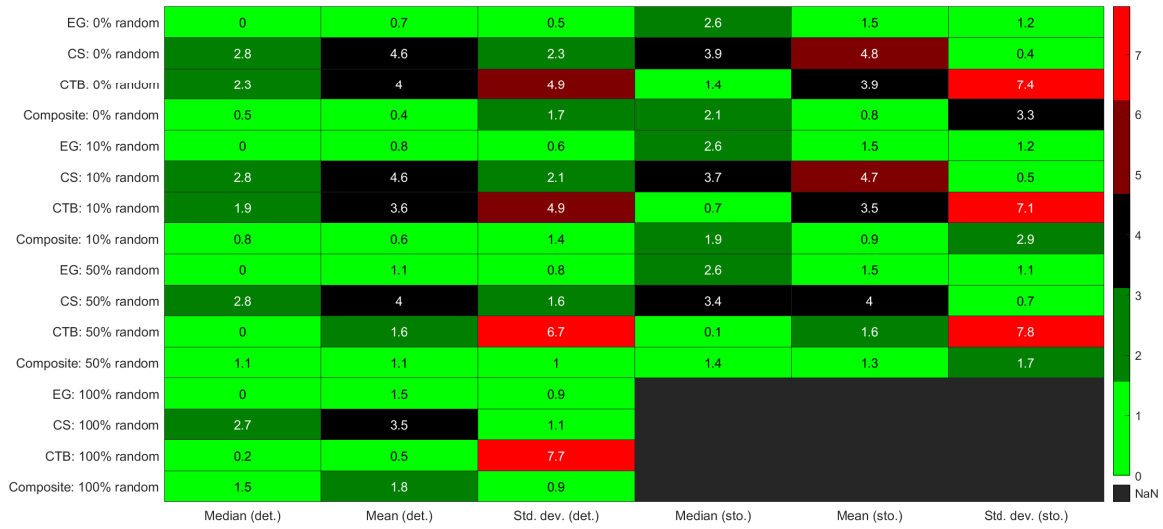
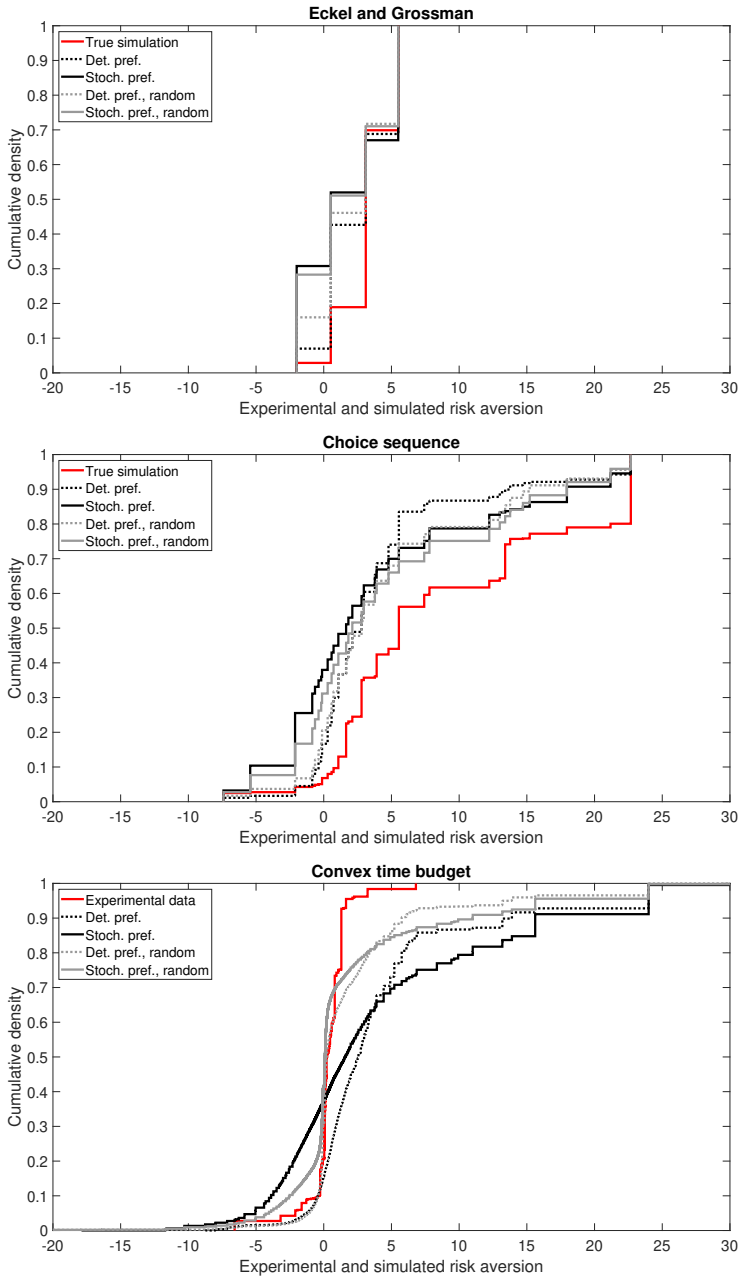


Figure 7: **Distance between simulated and experimental risk aversion values.** On the horizontal axis the mean, median, and standard deviation for deterministic preferences (columns 1-3) and stochastic preferences (columns 4-6). On the vertical axis the three elicitation methods, including the composite risk aversion measure (i.e., average of the three methods), grouped by the percentage of random choices.

Figure 8: **Comparison between simulated and experimental coefficients of risk aversion by task.** Number of virtual subjects 10,000.



4. Stability

To determine whether preferences are stable across methods, we use a preference stability measure based on overlapping intervals, similar to Holzmeister and Stefan (2021). We define choices in two tasks as "stable" if the implied parameter intervals overlap. That is, we define an indicator for each pairwise comparison of EG with CS and EG with CTB. As preference stability index, $stable_i$, we sum up these two binary indicators, implying a measure between 0 and 2 for each individual i. For example, if the point estimates of CS and CTB lie in the same EG interval, the stability index equals two.

As second stability measure, we define choices as "stable" if the ranking of an individual's risk aversion is consistent, similar to Frey et al. (2017). That is, we sort all risk aversion parameter values from low to high per elicitation task. The most risk seeking person receives ranking 1 and the most risk averse person receives ranking 1601, per method. To allow for a comparison with our virtually simulated subjects, we normalize the rankings on a scale between 0 and 1. Then, we create two stability measures for each individual i: (i) the average over the rankings across methods per individual, $rank_i^{\mu}$, and (ii) the standard deviation over the rankings across methods per individual, $rank_i^{\sigma}$.

For both measures, we study the stability of the experimentally observed risk preferences as well the simulated risk preferences. We study several types of simulated risk preferences: (i) deterministic and stochastic preferences, (ii) deterministic and stochastic preferences with 50% random choices, (iii) 100% random choices, and (iv) 100% random independent choices drawn from the experimentally observed distribution. The former three are identical to the simulations above, while the latter simulates the idea that subjects treat each of the tasks independently (Holzmeister and Stefan, 2021).

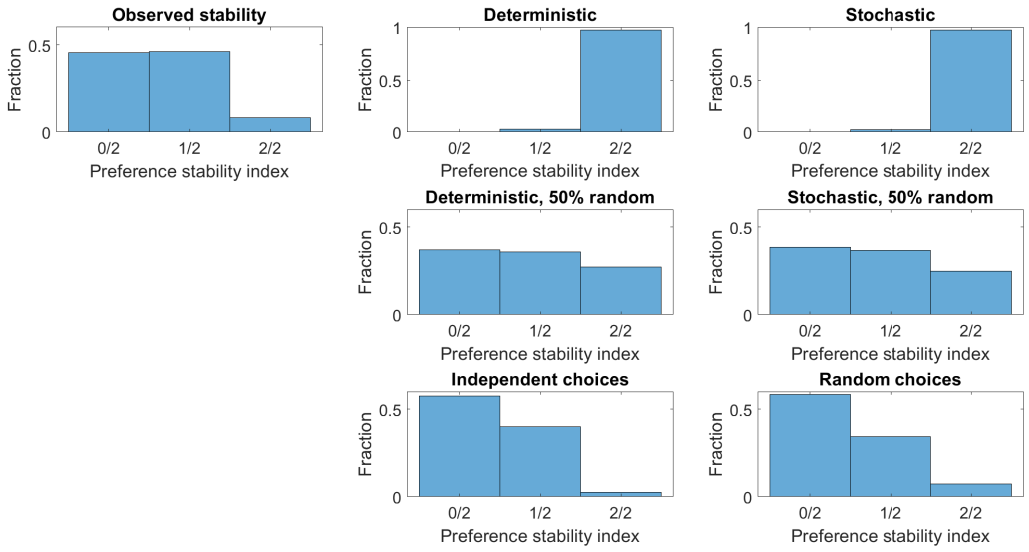


Figure 9: **Stability of experimental and simulated risk preferences: Overlapping intervals.**

A. Overlapping intervals

Figure 9 shows the distributions of the preference stability index observed in the experiment as well as the results of the simulations. About 45% of the observed experimental choices is unstable (i.e., 0/2) and about 10% is stable (i.e., 2/2). Eyeballing the distributions, the simulation outcomes with at least 50% of the subjects behaving randomly highlights considerable similarities with the observed experimental stability.

Table 5 presents the mean, median, and standard deviation of the distributions. At least 50% of random choices appears to be necessary to explain the observed experimental stability.

Table 5: **Stability of experimental and simulated risk preferences.** Number of virtual subjects 10,000.

Panel A: Overlapping intervals			
	Median	Mean	Std. dev.
Observed stability	1.00	0.63	0.63
Deterministic	2.00	1.97	0.17
Stochastic	2.00	1.97	0.17
Deterministic, 50% random	1.00	0.90	0.80
Stochastic, 50% random	1.00	0.86	0.78
Independent draws exp. data	0.00	0.45	0.55
Only random choices	0.00	0.49	0.63

Panel B: Rankings		
	Std. dev. over mean ranks	Mean over std. dev. of ranks
Observed stability	0.21	0.20
Deterministic	0.29	0.01
Stochastic	0.27	0.09
Deterministic, 50% random	0.21	0.21
Stochastic, 50% random	0.22	0.19
Independent draws exp. data	0.19	0.24
Only random choices	0.18	0.25

B. Rankings

Table 5, Panel B, shows the standard deviation of the averaged individual rankings, i.e., the standard deviation of $rank_i^{\mu}$, and the mean of the standard deviation of individual rankings, i.e., the average of $rank_i^{\sigma}$. A higher standard deviation of $rank_i^{\mu}$ implies more consistent choices, as this indicates more variability in the rankings. Namely, provided that there is sufficient resolution within and perfect consistency across measures, the resulting mean ranks would be uniformly distributed between 1 and 1601. Because EG and CS do not provide a resolution for 1601 distinct values, we also simulate perfect consistent behavior and completely random behavior. The results confirm our above findings: assuming that 50% of the population chooses randomly during the elicitation tasks matches observed stability exactly (0.21). The observed stability is closer to the benchmark of 'only random choices' rather than perfect consistency from 'deterministic'.

The average standard deviation of rankings yields similar results. Deterministic preferences yield, as expected, an average standard deviation of nearly zero. That is, individuals can realize an almost perfect ranking when behaving perfectly according to their true underlying risk aversion. With 50% of the subjects behaving randomly, the average standard deviations again are closest to the observed behavior.

Figure 10 represents these findings graphically. 50% random choice behavior (i.e., grey lines) lie almost exactly on the observed stability (i.e., red line).

Our results for both stability measures are almost identical for participants that took more than 3.5 minutes (15^{th} percentile) to complete the survey ($N = 590$, only available for those that did the 'classic questionnaire'). That is, participants that quickly answer the survey — and are perhaps more likely to answer randomly — are not driving our results.

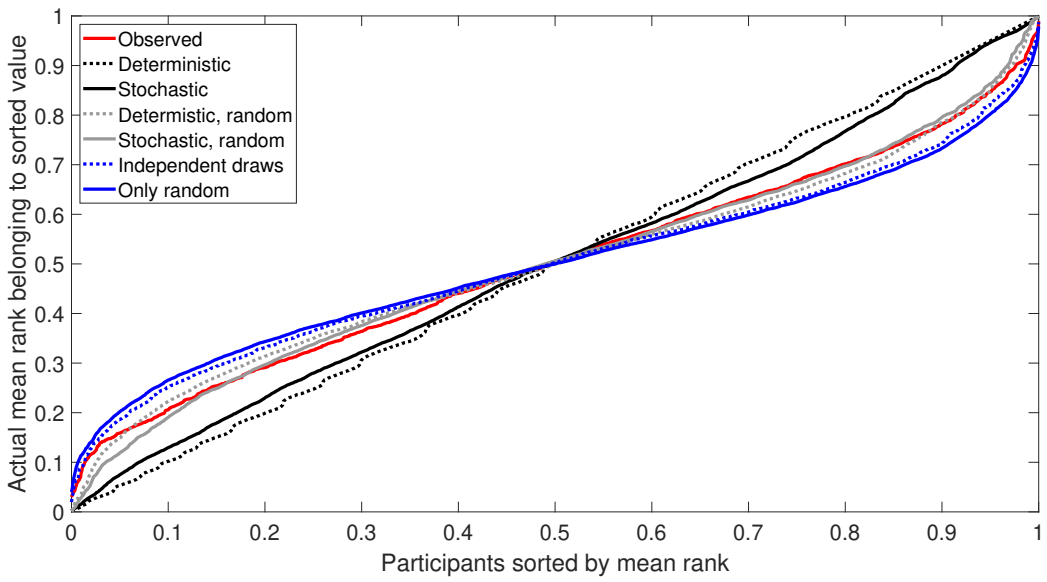


Figure 10: **Stability of experimental and simulated risk preferences: Rankings.**

C. Individual characteristics

In Table 6 we show mean individual characteristics for both stability measures as well as a statistical test for the difference between stable and unstable preferences.

White collar workers, high income earners, and young subjects have higher preference stability. We observe no difference in preference stability between males or females, singles or partners, and the game or classic questionnaire, where the latter is comforting. Interestingly, we observe more risk tolerant behavior at higher preference stability. This holds specifically for EG and CS, while for CTB the effect is reversed but also the difference in risk aversion magnitudes is rather small. Finally, it appears that high preference stability is accompanied by higher self-reported cognitive certainty, but statistically insignificant for the ranking measure.

Table 6: **Mean individual characteristics sorted by preference stability.** The second number per row variable indicates the number of observations. The p-values follow from a non-parametric Wilcoxon rank-sum test.

	Overlapping intervals				Rankings		
	Preference stability index			Diff. 0/2 - 0/2	Decile 1	Decile 10	Diff. 1 - 10
	0/2	1/2	2/2	p-value	(stable)	(unstable)	p-value
White collar worker	0.47	0.41	0.66	0.00	0.50	0.32	0.00
	583	606	119		142	130	
Income (Euros)	38894	37639	50571	0.00	42376	31703	0.00
	595	622	117		140	136	
Age (years)	59.84	59.5	55.89	0.00	57.64	63.08	0.00
	728	740	133		161	161	
Retired	0.33	0.35	0.26	0.09	0.33	0.47	0.01
	728	740	133		161	161	
Male	0.93	0.96	0.95	0.53	0.93	0.98	0.06
	728	740	133		161	161	
Partner	0.80	0.78	0.79	0.88	0.81	0.82	0.89
	728	740	133		161	161.00	
Game	0.56	0.57	0.60	0.35	0.53	0.57	0.58
	728	740	133		161	161	
Risk aversion EG	3.22	3.77	0.73	0.00	3.11	3.59	0.00
	728	740	133		161	161	
Risk aversion CS	9.19	9.97	1.47	0.00	7.59	14.96	0.00
	728	740	133		161	161	
Risk aversion CTB	0.28	0.34	0.6	0.23	0.36	-0.72	0.00
	728	740	133		161	161	
Risk aversion composite	4.23	4.69	0.93	0.00	3.69	5.94	0.00
	728	740	133		161	161	
Cognitive certainty EG	2.97	2.96	3.08	0.05	3.08	2.96	0.37
	728	740	133		161	161	
Cognitive certainty CS	2.84	2.87	2.96	0.04	2.96	2.87	0.40
	728	740	133		161	161	
Cognitive certainty CTB	2.89	2.91	3.02	0.03	3	2.95	0.78
	728	740	133		161	161	
Cognitive certainty total	8.70	8.74	9.06	0.02	9.04	8.78	0.75
	728	740	133		161	161	

5. Conclusion

We conduct a within-subject experiment with a representative population of 1601 pension fund participants that make realistic risky pension choices. We examine the heterogeneity in revealed risk preferences across three risk preference elicitation methods. In line with previous research (Pedroni et al., 2017), we find substantial variation in risk aversion estimates in terms of levels and rank order. We use two risk preference stability measures recently introduced in the literature (Holzmeister and Stefan, 2021; Frey et al., 2017) and find that, on average, at least 50% of the population makes random choices. That is, comparing the observed behavior to results from simulation exercises, we find that the observed stability of risk preferences across tasks is closest to behavior arising from at least 50% of the subjects making random choices per elicitation method. As such, our paper contributes to understanding the "risk elicitation puzzle" (Pedroni et al., 2017) in a representative sample: a potential explanation for the the risk elicitation puzzle is that a large fraction of subjects makes random choices.

References

Andreoni, J. and C. Sprenger (2012). "Estimating Time Preferences from Convex Budgets". In: *American Economic Review* 102.7, pp. 3333–3356.

Barsky, R., F. Juster, M. Kimball, and M. Shapiro (1997). "Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study". In: *Quarterly Journal of Economics* 112.2, pp. 537–579.

Cohen, J.D., K.M. Ericson, D. Laibson, and J.M. White (2020). "Measuring Time Preferences". In: *Journal of Economic Literature* 58.2, pp. 299–347.

Crosetto, P. and A. Filippin (2016). "A theoretical and experimental appraisal of four risk elicitation methods". In: *Experimental Economics* 19, pp. 613–641.

Eckel, C. and P. Grossman (2002). "Sex differences and statistical stereotyping in attitudes toward financial risk". In: *Evolution and Human Behavior* 23.4, pp. 281–295.

Eckel, C. C. and P. J. Grossman (2008). "Forecasting risk attitudes: An experimental study using actual and forecast gamble choices". In: *Journal of Economic Behavior & Organization* 68, pp. 1–17.

Frey, R., A. Pedroni, R. Mata, J. Rieskamp, and R. Hertwig (2017). "Risk preference shares the psychometric structure of major psychological traits". In: *Science Advances* 3.10, e1701381.

Goossens, J. and M. Knoef (2022). "COVID-19 Crisis: Do Extreme Events Affect Preferences and Trading Behavior?" In: *Working paper*.

Hackethal, A., M. Kirchler, C. Laudenbach, M. Razen, and A. Weber (2022). "On the role of monetary incentives in risk preference elicitation experiments". In: *Journal of Risk and Uncertainty*, pp. 637–648.

Holzmeister, F. and M. Stefan (2021). "The risk elicitation puzzle revisited: Across-methods (in)consistency?" In: *Experimental Economics* 24, pp. 593–616.

Pedroni, A., R. Frey, A. Bruhin, G. Dutilh, R. Hertwig, and J. Rieskamp (2017). "The risk elicitation puzzle". In: *Nature Human Behavior* 1, pp. 803–809.

Potters, J., A. Riedl, and P. Smeets (2016). "Towards a Practical and Scientifically Sound Tool for Measuring Time and Risk Preferences in Pension Savings Decisions". In: *Netspar Industry Paper* 59.

Appendix

Table 7: **Choice sequence risk aversion task.** Subjects choose a pension per question. Each question involves a choice between a risky and non-risky pension. A pension, either risky or non-risky, involves a 50/50 chance of a low or high payoff. A risky pension involves more risk, while a non-risky pension involves less risk. The impliedrisk aversion is based on the power utility function $U(x) = \frac{x^{1-\gamma}}{1-\gamma}$.

Question	Sequence	Risk aversion	Risky (R) High	Risky (R) Low	Non-risky (N) High	Non-risky (N) Low	Risk aversion after risky	Risk aversion after non-risky
1		14.50	3270	2050	2380	2080	2.49	10.07
2	N	-0.50	2640	1040	2050	1800	4.05	11.60
2	R	-0.50	2640	1040	2050	1800	-0.12	3.52
3	NN	11.60	3550	2210	2580	2260	3.98	16.72
3	NR	4.05	3550	2020	2580	2260	-0.43	13.68
3	RN	3.52	3550	1970	2580	2260	1.44	6.67
3	RR	-0.12	3550	1220	2580	2260	-1.90	1.65
4	NNN	16.72	2980	1880	2170	1900	14.11	20.53
4	NNR	3.98	2980	1690	2170	1900	1.64	8.11
4	NRN	13.68	2980	1870	2170	1900	6.96	17.71
4	NRR	-0.43	2980	820	2170	1900	-2.19	1.10
4	RNN	6.67	2980	1796	2170	1900	4.11	10.35
4	RNR	1.44	2980	1450	2170	1900	0.44	2.65
4	RRN	1.65	2980	1480	2170	1900	0.01	4.46
4	RRR	-1.90	2790	1070	2370	2100	-4.92	-0.29
5	NNNN	20.53	3810	2420	2780	2430	17.96	22.68
5	NNNR	14.11	3810	2400	2780	2430	12.22	15.21
5	NNRN	8.11	3810	2330	2780	2430	4.78	13.77
5	NNRR	1.64	3810	1900	2780	2430	0.58	2.96
5	NRNN	17.71	3810	2420	2780	2430	14.71	21.18
5	NRNR	6.96	3810	2300	2780	2430	1.82	13.00
5	NRRN	1.10	3810	1800	2780	2430	-0.19	3.78
5	NRRR	-2.19	3520	1770	3090	2730	-5.43	-0.38
5	RNNN	10.35	3810	2370	2780	2430	7.42	13.40
5	RNNR	4.11	3810	2170	2780	2430	2.79	5.55
5	RNRN	2.65	3810	2050	2780	2430	1.65	3.91
5	RNRR	0.44	3810	1590	2780	2430	-0.12	1.08
5	RRNN	4.46	3810	2200	2780	2430	2.11	7.81
5	RRNR	0.01	3810	1400	2780	2430	-0.65	0.72
5	RRRN	-0.29	3810	1190	2780	2430	-0.85	0.29
5	RRRR	-4.92	3410	1690	3190	2830	-7.41	-2.12

Table 8: **Eckel-Grossman risk aversion task.** Subjects choose a pension, all of which involve a 50/50 chance of a low or high payoff. The implied Coefficient of Relative Risk Aversion (CRRA) range is based on the power utility function $U(x) = \frac{x^{1-\gamma}}{1-\gamma}$. Each range is calculated by equalizing the gamble to its neighbors, and computing the value of γ that makes the individual indifferent in utility between each adjacent gamble.

Choice	Low payoff	High payoff	Exp. return	St. Dev.	Implied CRRA range
Pension 1	1970	2050	2010	57	$\gamma > 4.37$
Pension 2	1900	2150	2025	177	$1.84 < \gamma < 4.37$
Pension 3	1500	3000	2250	1061	$-0.80 < \gamma < 1.84$
Pension 4	1100	3200	2150	1485	$\gamma < -0.80$

Table 9: **Overview experimental design: Convex Time Budgets.** Choice sets in the Convex Time Budgets. t and k are front and end delays in years, and c_t and c_{t+k} are allocated amounts in Euros. $1+r$ is the implied gross interest rate. Annual r is the yearly interest rate in percent and calculated as $((1+r)^{1/k}-1)\times 100$.

Decision	t	k	c_t	c_{t+k}	$1+r$	Annual
1	1	10	10000	10000	1.00	0.00
2	1	10	10000	16000	1.60	4.81
3	1	10	10000	26000	2.60	10.03
4	1	5	10000	10000	1.00	0.00
5	1	5	10000	16000	1.60	9.86
6	1	5	10000	26000	2.60	21.06

Online appendix

Intro text (new page):

Thank you for wanting to participate in this bpfBOUW survey.

You are in charge of your pension and determine the risks!
We are going to look at pensions and you can choose. There is no right or wrong.

Your choices help to make your pension better and more personal.
In this study, assume that the prices of products and services will not change in the future.

Click on the arrow to start.

You now start with block 1. *(new page)*

The 1st question (new page):

Which pension do you choose?

With pension A you receive a total of €2,380 on your bank account every month if things are better than expected, or €2,080 if things are worse than expected. With pension B you will receive a total of €3,270 in your bank account every month if things are better than expected, or €2,050 if things are worse than expected. The chance that it will be better or worse is the same (50%), just like with heads or tails.

	Tails	Heads
A	€2.380	€2.080
B	€3.270	€2.050

The 2nd question (new page):
We ask you to choose pension A or B 4 more times. Note: the amounts change. Which pension do you choose?

[INSERT NEXT CHOICE TABLE, depends on previous choice]

The 3rd question (new page):
Which pension do you choose?

[INSERT NEXT CHOICE TABLE, depends on previous choice]

The 4th question (new page):
Which pension do you choose?

[INSERT NEXT CHOICE TABLE, depends on previous choice]

The 5th question (new page):
Which pension do you choose?

[INSERT NEXT CHOICE TABLE, depends on previous choice]

The 6th question (new page):
How sure are you about the answers you just gave? [Very insecure, insecure, sure, very sure]

You have completed block 1. You now start with block 2. *(new page)*

The 7th question (new page):
There are now 4 pensions to look at. You can only choose one. The pension amounts differ when things are better than expected and worse than expected. Which pension do you choose?

	Tails	Heads
A	€2.050	€1.970
B	€2.150	€1.900
C	€3.000	€1.500
D	€3.200	€1.100

The 8th question (new page):
How sure are you about the answers you just gave? [Very insecure, insecure, sure, very sure]

You have completed block 2. You now start with block 3. *(new page)*

The 9th question (new page):
You get coupons. You can buy everything with it in the following years. With voucher A you can spend €10,000 in 1 year and €0 in 10 years. With the other vouchers you will receive less in 1 year, but more later. Which voucher do you choose?

	Valid in 1 year	Valid in 10 years
Voucher A	€10.000	€0
Voucher B	€7.000	€3.000
Voucher C	€3.000	€7.000
Voucher D	€0	€10.000

The 10th question (new page):
We ask you to choose a voucher 2 more times. Note: the amounts change. Which voucher do you choose?

	Valid in 1 year	Valid in 10 years
Voucher A	€10.000	€0
Voucher B	€7.000	€4.800
Voucher C	€3.000	€11.200
Voucher D	€0	€16.000

The eleventh question (new page):
Which voucher do you choose?

	Valid in 1 year	Valid in 10 years
Voucher A	€10.000	€0
Voucher B	€7.000	€7.800
Voucher C	€3.000	€18.200
Voucher D	€0	€26.000

The 12th question (new page):
You will receive new vouchers. Now you can spend an amount **in 1 year and in 5 years**. Which voucher do you choose?

	Valid in 1 year	Valid in 5 years
Voucher A	€10.000	€0
Voucher B	€7.000	€3.000
Voucher C	€3.000	€7.000
Voucher D	€0	€10.000

The 13th question (new page):
We ask you to choose a voucher 2 more times. Note: the amounts change. Which voucher do you choose?

	Valid in 1 year	Valid in 5 years
Voucher A	€10.000	€0
Voucher B	€7.000	€4.800
Voucher C	€3.000	€11.200
Voucher D	€0	€16.000

The 14th question (new page):
Which voucher do you choose?

	Valid in 1 year	Valid in 5 years
Voucher A	€10.000	€0
Voucher B	€7.000	€7.800
Voucher C	€3.000	€18.200
Voucher D	€0	€26.000

The 15th question (new page):
How sure are you about the answers you just gave? [Very insecure, insecure, sure, very sure]

End (new page):

You are now done with this research. Thanks for filling in!

With your answers we can provide you with the best possible service in the future.

You can now close this screen.