

DISCUSSION PAPER SERIES

IZA DP No. 16343

Order Effects in Eliciting Preferences

Orestis Kopsacheilis
Sebastian J. Goerg

JULY 2023

DISCUSSION PAPER SERIES

IZA DP No. 16343

Order Effects in Eliciting Preferences

Orestis Kopsacheilis

Technical University of Munich

Sebastian J. Goerg

Technical University of Munich and IZA

JULY 2023

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Order Effects in Eliciting Preferences

Having an accurate account of preferences help governments design better policies for their citizens, organizations develop more efficient incentive schemes for their employees and adjust their product to better suit their clients' needs. The plethora of elicitation methods most commonly used can be broadly distinguished between methods that rely on people self-assessing and directly stating their preferences (qualitative) and methods that are indirectly inferring such preferences through choices in some task (quantitative). Alarming, the two approaches produce systematically different conclusions about preferences and, therefore, survey designers often include both quantitative and qualitative items. An important methodological question that is hitherto unaddressed is whether the order in which quantitative and qualitative items are encountered affects elicited preferences. We conduct three, pre-registered, studies with a total of 3,000 participants, where we elicit preferences about risk, time-discounting and altruism in variations of two conditions: 'Quantitative First' and 'Qualitative First'. We find significant and systematic order effects. Eliciting preferences through qualitative items first boosts inferred patience and altruism while using quantitative items first increases the cross-method correlation for risk and time preferences. We explore how monetary incentivization and introducing financial context modulates these results and discuss the implications of our findings in the context of nudging interventions as well as our understanding of the nature of preferences.

JEL Classification: C83, C91, D01, D91

Keywords: preferences, qualitative vs. quantitative measures, risk, altruism, patience

Corresponding author:

Orestis Kopsacheilis
Technical University of Munich
Arcisstraße 21
80333 München
Germany
E-mail: orestis.kopsacheilis@tum.de

1 Introduction

Social scientists study human preferences and resulting behavior through the lens of three fundamental trade-offs: risk vs. return, today vs. tomorrow and self vs. others. The ability to accurately elicit individuals’ preferences along these three dimensions has important consequences in a variety of applications and is, therefore, crucial to the design and the development of better institutions and organisations. For example, knowledge about risk preferences helps towards the provision of health-care plans that are better tuned to the individuals they set out to service. They also inform the management board about the risk tolerance of their typical customer profile, allowing them to customize their product to their customer needs. Accurate measurements of time preferences help provide better pension plans as well as design more appropriate dynamic incentive schemes. Moreover, well-calibrated accounts for people’s altruistic preferences informs the design of more appropriate redistribution policies and facilitates efficient personnel assortment into teams.

There are two distinct traditions in measuring these preferences. The first infers preferences by observing peoples’ choices that typically involve options expressed in monetary units. We refer to this as the ‘quantitative’ approach.¹ The second tradition asks instead people to self-assess and directly report their preference profile through some scale. We refer to this as the ‘qualitative’ approach.² By creating a choice environment that mimics decision-environments with real economic interest, quantitative measures are more likely to exhibit high external validity (though this supposition is not always straight-forward to demonstrate [Galizzi and Navarro-Martinez 2019](#)). Moreover, since quantitative questions can be monetarily incentivized in an incentive-compatible way, they avoid the ‘talk is cheap’ concern that is often associated with responses in qualitative items. On the other hand, qualitative questions are easier to explain to subjects as well as faster and cheaper to implement. They have, therefore, been the method of choice for practitioners or in large scale surveys. For example, survey data like the German Socio-Economic Panel or the National Longitudinal Study of Youth (US-based) employ exclusively qualitative items.

Ideally, the two traditions would converge to the same, or at least similar, conclusions. However, there is now sizeable evidence to the contrary (e.g. [Lönnqvist et al. 2015](#); [Pedroni et al. 2017](#); [Frey et al. 2017](#)). To overcome this conundrum, researchers often ‘hedge their bets’ by including both types of questions in their surveys. Additionally, modern approaches have tried to harness the best of the two worlds by combining quantitative with qualitative measures based on some estimated weight ([Falk et al., 2023](#)). This method has been used in one of the largest preferences elicitation from approximately 80,000 people in 76 countries ([Falk et al., 2018](#)).

A previously unaddressed question with potentially important methodological implications is whether the order in which the two types of elicitation methods are administered affects responses in a systematic way. Order effects occur when prior experience with one task affects behavior in a subsequent task and have been shown to be pertinent in the elicitation of preferences ([Harrison et al., 2005](#)). We conduct three, pre-registered studies with a total of 3,000 participants, where we elicit preferences about risk, time-discounting and altruism in variations of two conditions: ‘*Quantitative First*’ and ‘*Qualitative First*’. Echoing past literature, we find consistent evidence across all three studies that elicited preferences differ significantly across the two methods. According to measures

¹The terms ‘behavioral’ or ‘task’ are sometimes also used in this literature to describe what we refer here as ‘quantitative’.

²The terms ‘self-reports’ or ‘ask’ are also used in this literature to describe what we refer here as ‘qualitative’.

from qualitative, self-reports, people are more risk tolerant, patient and altruistic compared to those obtained from quantitative, behavioral tasks. What is more relevant for our investigation, however, is that we observe systematic order effects.

Specifically, in Study 1 we use the survey module as described in [Falk et al. \(2023\)](#). The preference module uses a quantitative and a qualitative item for each preference dimension. However, unlike the preference module that presents all qualitative items first and qualitative ones second, we randomize the order of these items for each preference dimension.³ We observe significant order effects of magnitude and consistency. With respect to magnitude, eliciting qualitative measures first increases inferred patience and altruism. With respect to consistency, eliciting quantitative measures first increases the cross-method correlation across methods for risk and time. In Study 2, we introduce incentive compatible versions of quantitative measures - the preference module uses hypothetical monetary rewards - and observe that such incentives inoculate quantitative measures from magnitude order effects. Order effects persist in qualitative measures as eliciting quantitative measures first reduces inferred patience and altruism in the qualitative measures. Moreover, eliciting quantitative measures first increases the cross-method correlation for risk and time. We hypothesize that (part of) the reason why we observe such persistent consistency order-effects is related to context mis-alignment. Quantitative questions are framed in monetary terms which is prompting a financial decision making context to most subjects whereas qualitative questions are framed abstractly, prompting differing contexts to different people. We test this hypothesis in Study 3 by framing qualitative items in financial context and verify that consistency order effects disappear in time-discounting preferences but are still (weakly) significant in risk.

Our study provides useful methodological insights. Our results not only point to the existence of significant order-effects in preference elicitations, they also provide a road-map on how to navigate and even harness them. Specifically, we find strong evidence that using incentive compatible payment methods inoculates quantitative measures from magnitude order effects. The item eliciting preferences for altruism is an exception to this as actual charity contributions are significantly higher when qualitative measures are elicited before. This result parallels with research on moral nudges ([Capraro et al., 2019](#)) and could be hinting towards promising paths for harnessing order effects for philanthropic goals. Our results also contribute to the growing evidence that preferences are context dependent. Our finding that apparent preference incongruence is (at least partly) driven by context mis-alignment, suggests that the elicitation puzzle – although present – may not be as severe as previously thought. A methodological implication is that it is important to frame qualitative items in the specific context of interest (e.g. financial decisions). Although the literature has readily available context-adjusted versions for risk ([Dohmen et al., 2012](#)) that are also widely used in surveys (e.g. German Socio-Economic Panel), to the best of our knowledge, we provide the first such formulations for time-discounting and altruism.

³It is worth pointing out that in our implementation, the quantitative and qualitative tasks were encountered back to back for every preference dimension, while in ([Falk et al., 2023](#)), all qualitative tasks for all preference dimensions were encountered before the quantitative ones. To the extent that distance between two similar tasks mitigates the spill over effect of one task towards the other, the size of the order effects we record in this paper should be seen as a ‘ceiling’.

2 Methods

Measures

We adapt the preference module, as described in [Falk et al. \(2023\)](#), and test for order effects between quantitative and qualitative items as well as how these are modulated by monetary incentivization and context in three studies. We focus on three preference dimensions: risk, time-discounting and altruism. Each preference dimension corresponds to one item and each item consists of a number of questions. Specifically, the quantitative item for altruism as well as all qualitative items consist of 1 question, while the quantitative item for time-discounting entails 25 and that for risk uses 31 questions. Unlike [Falk et al. \(2023\)](#) who administer the preference module in-person using pen and paper, we distribute the survey in an online, computerized setting. Subjects see only one question at a time and once an answer is submitted they can no longer go back and change it.⁴ Table 1 provides an overview of the preference module we adopt in this paper. For more details regarding the interface of each item, see Appendix A.

The quantitative item for eliciting risk preferences consists of a multiple price list. Participants make 31 choices between a lottery that remains constant and a safe option offering a certain amount that ranges from the highest to the lowest amount offered by the lottery. The lottery’s highest outcome is 300 and the lowest is 0 monetary units. This type of task is commonly used in the literature to infer risk preferences (e.g., [Holt and Laury 2002](#)). Switching from the safe amount to the lottery ‘late’ (i.e. when the safe amount is closer to 0) reveals higher risk-aversion (lower risk tolerance). This is consistent with Expected Utility Theory ([Bernoulli, 2011](#)) whereby risk averse people are willing to forego some return in order to avoid variance.⁵

The quantitative item for time-discounting also uses a multiple price list. In this case, the list consists of 25 choices between a payment today and a payment in 12 months. The payment today is fixed at 100 monetary units while that in 12 months ranges from 100 to 185 monetary units. Switching from the immediate payment to that in 12 months from now ‘early’ (i.e. where the delayed payment is close to 100) is associated with higher degrees of patience. In the context of the classical Discounted Utility model ([Samuelson, 1937](#)), this would correspond to a discounting factor closer to 1.⁶

Lastly, the quantitative item for altruism elicits the extent to which people are willing to give up money in order to improve someone else’s material payoff or well-being. The higher (lower) the proportion of someone’s own endowment that is being allocated to another party, the more altruistic (selfish) the individual is deemed to be. In the preference module this is implemented through a simple scale, with one end indicating the entire allocation to oneself while the other end to charity. In experimental economics literature, altruistic behavior has most commonly been studied through the dictator game where a player decides how much of an endowment to keep for themselves and how much they want to transfer to a second party who has a passive role ([Forsythe et al., 1994](#)). However, variations such as the one we implement here, where the second

⁴[Falk et al. \(2023\)](#) additionally elicit preferences for trust as well as positive and negative reciprocity.

⁵There is, however, a wealth of literature providing additional nuances to this fundamental trade-off. Prominent examples include Cumulative Prospect Theory ([Tversky and Kahneman, 1992](#)), Regret Theory ([Loomes and Sugden, 1982](#)), but see also [Starmers \(2000\)](#) for a more extensive overview.

⁶For models offering deeper behavioral insights on time preferences, see [Laibson \(1997\)](#) but also [Frederick et al. \(2002\)](#) for a (critical) review of this literature.

party is replaced with a non-profit institution outside the laboratory, are also common (Eckel and Grossman, 1996). Irrespective of the task used to elicit charitable giving, the consensus is that people exhibit a preference for giving; a conclusion running against the neoclassical account of a selfish agent who only cares about maximizing individual payoff.

For the qualitative items we always use Likert scales that allow for different levels of agreement to a certain statement. The statements can be seen on the third column of Table 1, while the degrees of agreement range from 0 to 10. The text associated with ‘0’ is: ‘Completely unwilling to take risks’, ‘Completely unwilling to give up something today’ and ‘Completely unwilling to share’ for risk, time and altruism preferences respectively. That for ‘10’ is identical, except that statements begin with ‘Very willing’. The formulations of these self-reports are based on items in existing surveys, like the German Socio-Economic Panel Study (SOEP), the National Longitudinal Study of Youth (NLSY) as well as previous research (e.g., Weber et al. 2002; Perugini et al. 2003).

The final measures for all six items are standardized between 0 and 1. For the qualitative measures we take the degree of agreement and divide with the range of the scale (‘10’ for all three survey items). Similarly, for the quantitative measure of altruism, we divide the amount contributed by the range of that scale (‘1,000’ monetary units in Studies 1 and 3; ‘100’ monetary units in Study 2). For the quantitative measures for risk and time-discounting preferences we take the switching point and divide it with the number of available items (‘31’ for risk; ‘25’ for time). In every case, a standardized score of ‘1’ corresponds to extreme risk tolerance, patience and altruism (and vice versa for ‘0’).

We use these standardized measures to examine the presence and size of order effects. We distinguish between two types of order effects: magnitude and consistency. Magnitude order effects are measured in the difference of average risk tolerance, patience or altruism between the two treatments for each measure. Consistency order effects are measured in the difference of the correlation coefficient between preferences inferred via quantitative and qualitative items. We use Wilcoxon rank-sum and Fisher’s z tests to test for the statistical significance of treatment differences in magnitude and consistency order effects respectively. Our analysis plan has been pre-registered.⁷

Studies

We conduct three studies, each with 1,000 subjects that we recruit online through Prolific Academic. Study 1 implements the preference module as described in the previous section and as summarised in Table 1.

Study 2 introduces incentive-compatibility for quantitative items. We implement a 1 to 10 currency exchange rate between the monetary units that were used in Study 1. Under this rate, the high outcome in the lottery is worth £30. For altruism, we shrink the initial endowment down to 100 monetary units, which now corresponds to £10. There are no differences in the qualitative items compared to Study 1.

In Study 3 we introduce financial context in the formulation of the questions in qualitative items. The statements on which subjects are asked to express agreement read as follows. **Risk:** ‘*How do you evaluate your attitude towards risk regarding financial investments?*’; **Time-discounting:** ‘*In comparison to others, are you a person who is willing to save money today in order to benefit from*

⁷The pre-registration can be accessed at https://aspredicted.org/44K_1R5

Table 1: The preference module (Falk et al., 2023) as applied in Study 1

Preference	Quantitative item	Qualitative item
Risk Taking	Multiple price list of 31 [hypothetical] choices between a 2-outcome lottery and a monetary amount offered with certainty	How do you see yourself: Are you a person who is [generally] willing to take risks, or do you try to avoid taking risks?
Time Discounting	Multiple price list of 25 [hypothetical] choices between an early payment ‘to-day’ and a delayed payment ‘in 12 months’	In comparison to others, are you a person who is [generally] willing to give up something today in order to benefit from that in the future?
Altruism	A [hypothetical] allocation of money to charity	How do you assess your willingness to share with others without expecting anything in return when it comes to charity?

Note. We implement an incentive compatible compensation scheme for Study 2 and, therefore, change the hypothetical wording accordingly. Moreover, Study 2’s altruism question features a drop-down menu with a list of well-known charities, including an option to specify a charity of their own if none of the already provided options suits them. For Study 3, we change the term ‘generally’ from the qualitative items and replace it with terms that reflect ‘financial decisions’.

*the financial gains of this investment in the future or are you not willing to do so ?’; **Altruism:** ‘How do you assess your willingness to share with others without expecting anything in return when it comes to donating money to charity?’* The quantitative measures are presented with hypothetical incentives, just as in Study 1. Table 2 summarises the key features of each study.

Table 2: Overview of studies

Study	Quantitative item	Qualitative item
Study 1	Hypothetical incentives	Wording in a general context
Study 2	Incentive compatible scheme	Wording in a general context
Study 3	Hypothetical incentives	Wording in financial context

The median completion time for sessions in Studies 1 and 3 is four minutes while for Study 2, it is five minutes. Each participant receives a flat £1 for their participation in Study 1 and £0.8 in Studies 2 and 3. Additionally, in Study 2, one out of twenty participants have one of their answers in the quantitative questions (across all three items) played out for real which earned those subjects an additional £16 - on average.

Treatments

In every study we randomly assign subjects in one of two treatments: ‘*Quantitative First*’ (Quant-First) or ‘*Qualitative First*’ (Qual-First). The only difference between the two treatments is that in the *Quantitative First* (*Qualitative First*) treatment subjects saw the quantitative (qualitative)

item before the qualitative (quantitative) one. This was the case for every preference dimension. The order in which subjects encountered each preference dimension was randomised.

Table 3 summarises some basic demographic information. With the exception of gender, which appears to be imbalanced in favor of females in Study 1, there are no striking differences across treatments and studies.

Table 3: Demographics

Study	Treatment	Age	Female	Education	Income
Study1	Qual-First	3.924	0.665	6.282	2.438
		(1.363)	(0.473)	(1.436)	(1.116)
Study1	Quant-First	3.875	0.612	6.243	2.506
		(1.286)	(0.488)	(1.477)	(1.126)
Study2	Qual-First	4.209	0.546	6.274	2.523
		(1.376)	(0.498)	(1.479)	(1.106)
Study2	Quant-First	4.260	0.520	6.246	2.500
		(1.42)	(0.5)	(1.493)	(1.101)
Study3	Qual-First	4.176	0.545	6.090	2.473
		(1.423)	(0.498)	(1.489)	(1.114)
Study3	Quant-First	4.182	0.51	6.208	2.544
		(1.405)	(0.500)	(1.488)	(1.089)

Note. Standard deviations in parentheses. ‘Qual-First’ (‘Quant-First’), refers to the treatment in which subjects encounter the qualitative (quantitative) item before the quantitative (qualitative) one for every preference dimension.

3 Results

Figure 1 provides a visual impression of our data across preference dimensions, treatments and studies. Preliminary visual inspection reveals noticeable differences *between* the two elicitation methods: quantitative measures (y-axis) and qualitative measures (x-axis). Most notably, average measures (across treatments) for risk, time-discounting and altruism are higher in qualitative items than in quantitative ones. This impression is consistent across all three studies and statistically significant: Mann-Whitney tests always reject the hypothesis that the two measures are equal ($p\text{-value} < 0.05$, for every preference dimension and every study).⁸ The focus of this paper, however, is the comparison of these measures *within* each elicitation method and across treatments. It is through these comparisons that we can focus on potential order effects.

Focusing on Study 1 (top row), we can see from the density plots that distributions of preference measures within each elicitation method is more similar than across methods. Nonetheless, the distribution of individual measures in the *Qualitative First* treatment is shifted towards higher values. This is particularly evident for time and altruism, foreshadowing the presence of significant magnitude order effects.

The constant and slope of the plotted linear models derive from Ordinary Least Squares regressions of quantitative on qualitative measures. The higher slope in *Quantitative First* relative to *Qualitative First* in risk and time preferences points towards the conclusion that encountering quantitative items before qualitative ones increases the cross-method correlation in these preference dimensions. One exception to this is with respect to altruism, where the opposite appears to be the case. Nonetheless, the difference in slope between the two treatments is smaller and as pointed out in a later stage of the analysis, not statistically significant.

These impressions are largely similar in Studies 2 and 3, albeit, with two notable exceptions. First, the mean of quantitative measures in Study 2 remains unchanged across treatments for risk and time preferences, suggesting that incentive compatibility mitigates magnitude order effects that occur when qualitative items precede quantitative ones. Second, the slope difference, evident in Studies 1 and 2 in time preferences, disappear in Study 3. This suggests that (at least part of) the apparent dissonance between the two elicitation methods is due to differences in context.

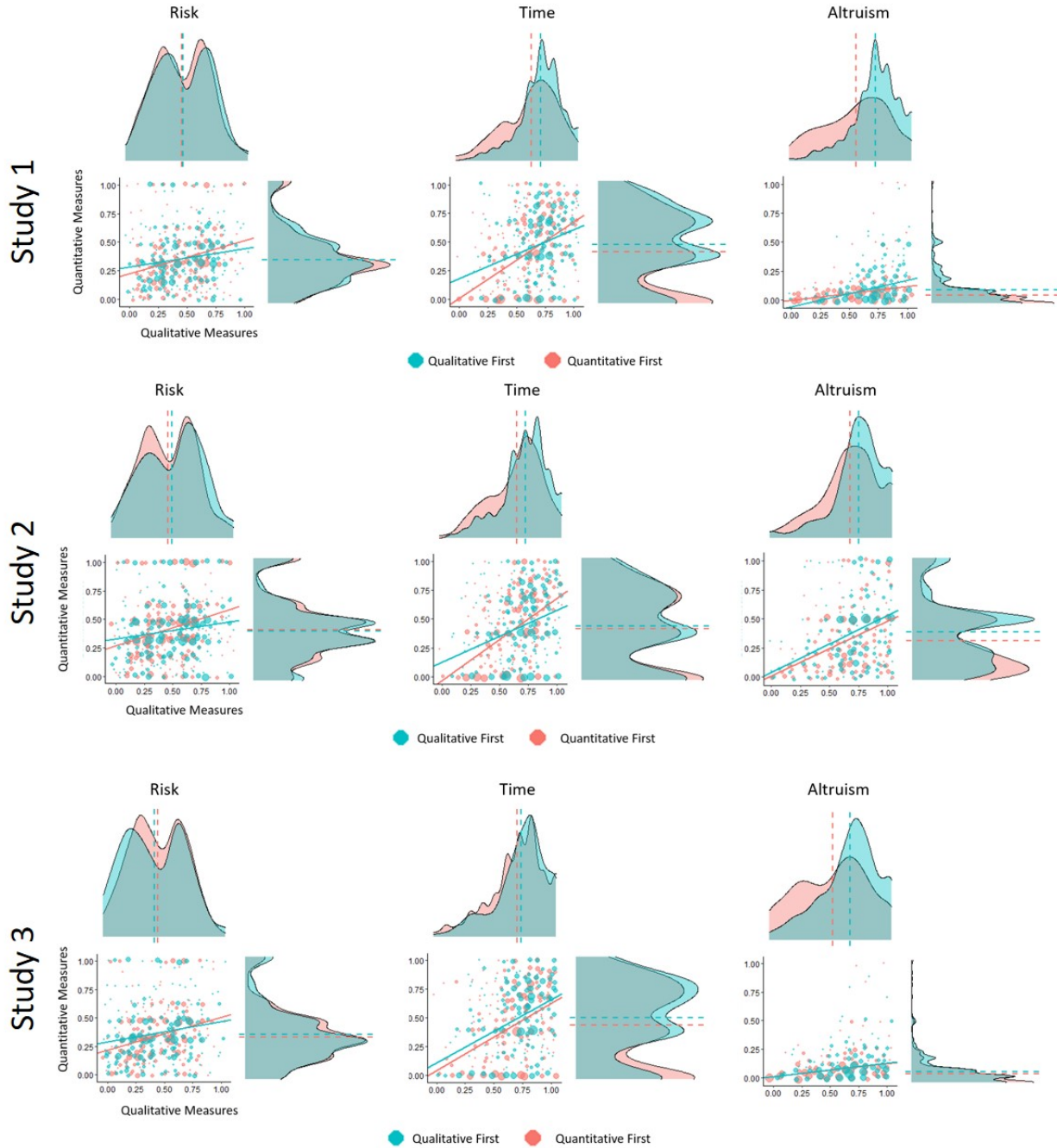
In what follows we analyze those visual impressions, together with the corresponding test-statistics, by distinguishing between magnitude (3.1) and consistency (3.2) order effects.

3.1 Magnitude order effects

We find significant evidence that the tendency of qualitative items to overstate – relatively to quantitative ones – people’s patience and altruism has further consequences. The averages for each measure as well as the p-values derived from the Mann-Whitney (MW) tests are detailed in Table 4. In Study 1, where quantitative items are hypothetically incentivized and quantitative items are framed in a general context, subjects’ behavior is consistent with higher levels of patience and altruism in *Qualitative First* compared to *Quantitative First* (time: $\mu_{Qn} = 0.416$, $\mu_{Ql} = 0.478$, $p =$

⁸Using t-tests instead of these non-parametric tests does not affect our results.

Figure 1: Scatterplots of quantitative and qualitative measures across treatments and studies, with partial density plots



Note. The slope and the constant of the plotted linear models derive from the Ordinary Least Squares regression of the quantitative on the qualitative measure. Dotted lines represent averages.

0.001; altruism: $\mu_{Qn} = 0.063$, $\mu_{Ql} = 0.104$, $p < 0.001$).⁹ There is no significant difference for risk ($\mu_{Qn} = 0.354$, $\mu_{Ql} = 0.355$, $p = 0.592$). The same tendency is observed for measures obtained from qualitative items. Subjects in *Qualitative First* self-report higher degrees of patience ($\mu_{Qn} = 0.615$, $\mu_{Ql} = 0.695$, $p < 0.001$) and altruism ($\mu_{Qn} = 0.544$, $\mu_{Ql} = 0.705$, $p < 0.001$) compared to *Quantitative First*. Again, risk preferences are impervious to such magnitude order effects ($\mu_{Qn} = 0.471$, $\mu_{Ql} = 0.455$, $p = 0.249$).

Table 4: Magnitude order effects

Quantitative measures				Qualitative measures		
Study 1	Hypothetical incentives			General context		
	Quant-First	Qual-First	p-value	Quant-First	Qual-First	p-value
Risk	0.354 (0.235)	0.355 (0.222)	0.592	0.471 (0.230)	0.455 (0.221)	0.249
Time	0.416 (0.311)	0.478 (0.291)	0.001	0.615 (0.227)	0.695 (0.182)	0.000
Altruism	0.063 (0.112)	0.104 (0.134)	0.000	0.544 (0.268)	0.705 (0.199)	0.000
Study 2	Incentive compatible			General context		
Risk	0.415 (0.260)	0.404 (0.256)	0.549	0.463 (0.226)	0.497 (0.249)	0.016
Time	0.418 (0.318)	0.44 (0.312)	0.274	0.633 (0.215)	0.701 (0.194)	0.000
Altruism	0.318 (0.276)	0.393 (0.283)	0.000	0.653 (0.238)	0.724 (0.211)	0.000
Study 3	Hypothetical incentives			Financial context		
Risk	0.345 (0.226)	0.364 (0.249)	0.577	0.443 (0.222)	0.417 (0.248)	0.057
Time	0.440 (0.323)	0.500 (0.306)	0.004	0.683 (0.225)	0.717 (0.212)	0.012
Altruism	0.067 (0.105)	0.085 (0.109)	0.000	0.519 (0.287)	0.657 (0.239)	0.000

Note. Standard deviations in parentheses. ‘Quant-First’, refers to the treatment in which subjects encounter the quantitative item before the qualitative one. ‘Qual-First’ refers to the treatment where the opposite is the case.

In Study 2 we introduce an incentive compatible scheme for quantitative items. Just as in Study 1, we observe no magnitude order effects for risk preferences ($\mu_{Qn} = 0.415$, $\mu_{Ql} = 0.404$, $p = 0.549$). In addition, we see that monetary incentivization in an incentive compatible way inoculates quantitative measures from magnitude order effects in timediscounting preferences ($\mu_{Qn} = 0.418$, $\mu_{Ql} = 0.440$; $p = 0.274$), which were present in Study 1. However, people are still behaving more altruistic in *Qualitative First* compared to *Quantitative First* ($\mu_{Qn} = 0.318$, $\mu_{Ql} = 0.393$; $p < 0.001$). This translated into an average increase of 24% of monetary contributions to charities (from £3.18 to £3.93). Moreover, magnitude order effects are still present in qualitative items. Subjects in *Quantitative First* self-assess to be more risk averse, impatient and selfish compared to those in

⁹The abbreviations μ_{Qn} and μ_{Ql} refer to the means of the measures in **Quantitative First** and **Qualitative First**, respectively.

Qualitative First. These effects are consistently statistically significant at 5% level in MW tests (risk: $\mu_{Qn} = 0.463, \mu_{Ql} = 0.497; p = 0.016$; time: $\mu_{Qn} = 0.633, \mu_{Ql} = 0.701, p < 0.001$; altruism: $\mu_{Qn} = 0.653, \mu_{Ql} = 0.724, p < 0.001$).

The results from Study 3 further reassure us that incentive compatibility is the key driver behind this inoculation effect. When we revert to hypothetical incentives for quantitative items, magnitude order effect reappear in time preferences for quantitative measures ($\mu_{Qn} = 0.440, \mu_{Ql} = 0.500; p = 0.004$). This suggests that aligning the context so that it prompts financial decision making across elicitation methods, does not affect magnitude order effects (but does impact consistency order effects as we discuss in the next section). Just like in all of our three studies, there are no magnitude order effects in quantitative measures for risk ($\mu_{Qn} = 0.345, \mu_{Ql} = 0.364; p = 0.577$) while there are statistically significant ones for altruism ($\mu_{Qn} = 0.067, \mu_{Ql} = 0.085; p < 0.001$). We also observe the same pattern in qualitative measures as in Studies 1 and 2. Specifically, subjects in *Quantitative First* self-report higher levels of impatience and selfishness compared to those in *Qualitative First*. These effects are consistently statistically significant at 5% level in MW (time: $\mu_{Qn} = 0.683, \mu_{Ql} = 0.717, p = 0.014$; altruism: $\mu_{Qn} = 0.519, \mu_{Ql} = 0.657, p < 0.001$). Risk preferences are an exception to this, where people’s self-assessment is consistent with higher risk tolerance in *Quantitative First* rather than in *Qualitative First*. Nonetheless, this is significant only at 10% (risk: $\mu_{Qn} = 0.443, \mu_{Ql} = 0.417; p = 0.057$).

Although we reserve all our nominal (and statistical) comparisons within each study, it is worth noting that average measures are remarkably consistent across studies for each preference dimension and each elicitation method. One exception to this is with respect to altruism where the quantitative measure in Study 2 is strikingly higher than that in Study 1 or 3 (from an average of 0.084 and 0.076 in Study 1 and 3 respectively, to an average of 0.356 in Study 2). This is because the scale was readjusted from 1-1000 (Study 1 and Study 3) to 1-100 (Study 2).

3.2 Consistency order effects

Table 5 summarises the Pearson correlations between quantitative and qualitative measures as well as the p-values obtained from Fisher z-scores.¹⁰

In Study 1 we observe that the cross-method correlation between quantitative and qualitative measures increases significantly in the *Quantitative First* treatment compared to the *Qualitative First* one. This is true for both risk ($r_1 = 0.274, r_2 = 0.165; p = 0.070$) and patience ($r_1 = 0.485, r_2 = 0.268; p < 0.01$) but not for altruism ($r_1 = 0.284, r_2 = 0.344; p = 0.295$). This pattern remains the same in Study 2, with the introduction of monetary incentives (risk: $r_1 = 0.276, r_2 = 0.144; p = 0.028$; time: $r_1 = 0.486, r_2 = 0.282; p < 0.001$; altruism: $r_1 = 0.405, r_2 = 0.266; p = 0.467$). However, we find that introducing financial context in qualitative items mitigates this asymmetry. Specifically, although the cross-method correlation is still (weakly) significantly higher in *Quantitative First* compared to *Qualitative First* ($r_1 = 0.285, r_2 = 0.181; p = 0.083$), the cross-method difference in correlation for time preferences is no longer statistically significant ($r_1 = 0.409, r_2 = 0.382; p = 0.607$). Just like in Studies 1 and 2, the cross-method correlation for altruism does not differ significantly across treatments ($r_1 = 0.306, r_2 = 0.264; p = 0.607$).

¹⁰Repeating these tests after transforming these measures into ranks relative to the distribution of the scores within each treatment, rather than absolute ones, does not affect our results.

Table 5: Consistency order effects

<i>Study 1</i>	<i>Quant: Hypothetical incentives/ Qual: General context</i>		
	Quant-First	Qual-First	p-value
Risk	0.274	0.165	0.070
Time	0.485	0.268	0.000
Altruism	0.284	0.344	0.295
<i>Study 2</i>	<i>Quant: Incentive compatible/ Qual: General context</i>		
Risk	0.276	0.144	0.028
Time	0.486	0.282	0.000
Altruism	0.405	0.366	0.467
<i>Study 3</i>	<i>Quant: Hypothetical incentives/ Qual: Financial context</i>		
Risk	0.285	0.181	0.0831
Time	0.409	0.382	0.607
Altruism	0.306	0.264	0.471

Note. ‘Qual-First’, refers to the treatment in which subjects encounter the qualitative item before the quantitative one for every preference dimension. ‘Quant-First’ refers to the treatment where the opposite is the case.

4 Discussion and Conclusion

Eliciting preferences accurately is of vital importance for social scientists and practitioners alike. Being able to gauge the risk tolerance, patience and altruism of individuals provides useful insights for informing public policy as well as managing organisations. According to one school of thought, the simplest and best way to learn about such preferences is by directly asking people to introspect and report the intensity of their preferences (e.g. how risk tolerant they are) on a Likert scale. A second approach, however, advocates for inferring such preferences through observed choices that involve (sometimes hypothetical) financial incentives. Alarming, there is now considerable evidence pointing to (apparent) systematic inconsistencies between these two traditions: *qualitative self-reports* and *quantitative behavioral tasks*. Therefore, it is becoming increasingly common to include both methods in surveys, sometimes in order to compare (e.g. Lönnqvist et al. 2015) and other times in order to harness the best of the two worlds by combining them into a hybrid score (Falk et al., 2018, 2023).

An important open question is whether there are systematic order effects between these qualitative and quantitative measures. In tackling this question, our study provides useful methodological insights. Our results not only point to the existence of significant order-effects in preference elicitation, they also provide a road-map on how to navigate as well as, potentially, harnessing them.

We conduct three, online studies, with a total of 3,000 subjects, where we test for order effects in eliciting preferences about risk, time-discounting and altruism. We do this by comparing differences between two treatments: *Quantitative First*, where quantitative items precede qualitative ones for every preference dimension and *Qualitative First* where the opposite is the case. Across all three

studies, we find consistent evidence suggesting that order effects are present and affect both: the level of the elicited measures (magnitude) as well as their cross-method correlation (consistency).

With respect to magnitude order effects, we find that when incentives are hypothetical (Studies 1 and 3), going through the qualitative self-assessment first boosts inferred patience and altruism in both quantitative and qualitative measures, relatively to when people go through the quantitative item first. Incentive compatibility (Study 2) inoculates quantitative measures elicited in time-discounting preferences but not with respect to altruism. Qualitative measures, on the other hand, remain susceptible to such magnitude order effects across all three studies.

One interpretation of the magnitude order effects relates to self-image concerns. People prefer to see themselves as relatively risk tolerant, patient and altruistic but might be less prone to ‘live up’ to this image when it is linked to a costly action. The logic behind this line of argument can be seen as an extension of the ‘talk is cheap’ criticism that is often ascribed to measurements elicited through qualitative self-reports. Interestingly, we find evidence for this even when the incentives for quantitative items are hypothetical (Studies 1 and 3). Throughout all of our three studies, subjects self-reported higher levels of risk tolerance, patience and altruism compared to what their behavior suggested in quantitative tasks. It is likely, that (at least part) of the magnitude order effects that we observe across the three studies is due to spill-over effects: the score of the desirable attribute is inflated when the qualitative question is encountered first and deflated when the quantitative question is encountered first instead. It is also important, however, to keep in mind that nominal discrepancies across elicitation methods need not be evidence of the methods’ incongruence. For example, someone who self-reports as extremely altruistic, e.g. ‘10/10’ in the corresponding qualitative question, might (legitimately) consider donating 1% of their endowment to be perfectly consistent with this self-assessment.

The ‘talk is cheap’ dictum is further corroborated by our finding that introducing an incentive-compatible scheme inoculates quantitative measures from such magnitude order effects - while leaving qualitative measures still susceptible to them. Interestingly, the quantitative measure for altruism is an exception to this - possibly because self-image concerns regarding altruism are stronger than for risk or time-discounting. Specifically, we observe a 24% increase in charitable donations when people report how altruistic they consider themselves to be before they are asked to donate compared to the alternative with the reverse order. This result is redolent of the findings related to moral nudges. [Capraro et al. \(2019\)](#) find that when they ask people what they think is the morally right thing to do, ensuing charitable giving goes up by about 44% compared to when they are asked to contribute without this moral assessment. We argue that the introspective nature of the self-reported, qualitative items evokes normative considerations about what is the ‘right’ thing to do. In this sense, qualitative measures tap into a similar mechanism as the one described in moral nudges. We believe that delving deeper into this mechanism is an avenue of promising future research with applications to charity-giving. For example, in our study we allow subjects to contribute up to £10. It would be interesting to see in a follow-up investigation how the scale of the amount donated modulates - if at all - these results.

To investigate consistency order effects we rely on cross-method correlations of quantitative and qualitative measures between treatments. At face value, the overall small to medium correlations we observe in Studies 1 and 2, across both treatments, echo the alarming sounds of past literature, pointing to fundamental incongruencies between the two elicitation traditions. Interestingly, we observe that eliciting quantitative measures before qualitative ones increases the cross method-

correlation significantly for risk and time preferences (but not altruism) throughout Studies 1 and 2 where qualitative questions are framed in abstract contexts (e.g. ‘are you a person who is **generally** willing to take risks?’). However, in Study 3, where we frame qualitative measures in financial context (e.g. ‘how do you evaluate your attitude towards risk regarding **financial investments**?’), the cross-method correlation in ‘Qualitative First’ is enhanced for time preferences (from 0.26 and 0.28 in Studies 1 and 2 to 0.36 in Study 3), rendering consistency order-effects no longer statistically significant. A similar tendency is observed for risk-preferences, but the consistency order effect is still (weakly) significant in that case.

These results corroborate our intuition that (at least part of) the reason behind this dissonance is due to the abstract context in which qualitative items are commonly presented to survey-takers. Abstract contexts have been shown to evoke different contexts for different people (Birnbaum, 1999). For example, asking how risk tolerant someone is ‘in general’ can prompt scenarios of financial decision making to some, while health decisions to others. In contrast, quantitative items typically require respondents to think in terms of financial trade-offs, imposing a context of financial decision making. To the extent that preferences are context dependent and someone’s willingness to take risks in health-related decisions differs from that in financial ones, eliciting qualitative measures first will lead to lower correlation scores. According to this conclusion, apparent incongruencies are (partly) mitigated when quantitative measures are elicited first since the context is now specified to be that of financial decision making.

Our results contribute to the growing evidence that preferences are largely context dependent. We identify two interesting implications from this. First, to the extent that the apparent incongruence between these two elicitation traditions was gauged using abstract frames in qualitative items, then the problem might have been exaggerated. Coupled with the finding by Holzmeister and Stefan (2021) that subjects are aware of the variation they exhibit across different elicitation methods, these results add important pieces to the ‘preference elicitation puzzle’.

Second, a methodological implication relates to survey designs. Specifically, it strongly suggests that quantitative items involving monetary trade-offs emphasize a context of financial decision making. If the researcher is interested in a different domain, e.g. health decisions, then it is important to explicitly frame the decision in those terms. Similarly, when qualitative items are framed in a general context, the researcher cannot be certain about the context respondents project onto the question. To this end, specifying the context of interest is pivotal. Although the literature has readily available context-adjusted versions for risk - for example Dohmen et al. (2012)’s questionnaire which is also used by German Socio-Economic panel - to the best of our knowledge, we provide the first such formulations of qualitative items for time-discounting and altruism.

References

- Bernoulli, D. (2011). Exposition of a new theory on the measurement of risk. In *The Kelly capital growth investment criterion: Theory and practice*, pages 11–24. World Scientific.
- Birnbaum, M. H. (1999). How to show that 9_i 221: Collect judgments in a between-subjects design. *Psychological Methods*, 4(3):243.
- Capraro, V., Jagfeld, G., Klein, R., Mul, M., and de Pol, I. v. (2019). Increasing altruistic and cooperative behaviour with simple moral nudges. *Scientific reports*, 9(1):1–11.
- Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2012). The intergenerational transmission of risk and trust attitudes. *The Review of Economic Studies*, 79(2):645–677.
- Eckel, C. C. and Grossman, P. J. (1996). Altruism in anonymous dictator games. *Games and economic behavior*, 16(2):181–191.
- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.
- Falk, A., Becker, A., Dohmen, T., Huffman, D., and Sunde, U. (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science*, 69(4):1935–1950.
- Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic behavior*, 6(3):347–369.
- Frederick, S., Loewenstein, G., and O’donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of economic literature*, 40(2):351–401.
- Frey, R., Pedroni, A., Mata, R., Rieskamp, J., and Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science advances*, 3(10):e1701381.
- Galizzi, M. M. and Navarro-Martinez, D. (2019). On the external validity of social preference games: a systematic lab-field study. *Management Science*, 65(3):976–1002.
- Harrison, G. W., Johnson, E., McInnes, M. M., and Rutström, E. E. (2005). Risk aversion and incentive effects: Comment. *American Economic Review*, 95(3):897–901.
- Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American economic review*, 92(5):1644–1655.
- Holzmeister, F. and Stefan, M. (2021). The risk elicitation puzzle revisited: Across-methods (in) consistency? *Experimental Economics*, 24(2):593–616.
- Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478.
- Lönnqvist, J.-E., Verkasalo, M., Walkowitz, G., and Wichardt, P. C. (2015). Measuring individual risk attitudes in the lab: Task or ask? an empirical comparison. *Journal of Economic Behavior & Organization*, 119:254–266.
- Loomes, G. and Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal*, 92(368):805–824.

- Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., and Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, 1(11):803–809.
- Perugini, M., Gallucci, M., Presaghi, F., and Ercolani, A. P. (2003). The personal norm of reciprocity. *European Journal of Personality*, 17(4):251–283.
- Samuelson, P. A. (1937). A note on measurement of utility. *The review of economic studies*, 4(2):155–161.
- Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of economic literature*, 38(2):332–382.
- Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5:297–323.
- Weber, E. U., Blais, A.-R., and Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of behavioral decision making*, 15(4):263–290.

A Instructions

A.1 Risk preferences

Figures A1a and A1b depict the basic interface for eliciting risk preferences - as implemented in Study 1. In Study 2, where we introduce an incentive compatible payment scheme, we stress that the monetary amounts are real and remove the hypothetical framing. Although we keep the numerical values of the outcomes the same, we remove the £symbol and specify that there is a 1 to 10 conversion rate to pounds so that 300 monetary units correspond to £30. In Study 3, we change the context from general to financial decision making. The text for the qualitative item for risk preferences in study 3 read as follows: ‘*How do you evaluate your attitude towards risk regarding financial investments?*’

Figure A1: Interface for eliciting risk preferences.

Please imagine the following situation: You can choose between a sure payment of a particular amount of money, or a draw, where you would have an equal chance of getting 300 pounds or getting nothing. We will present to you different situations - each row represents a different situation. Option A (the draw) is the same in all situations. Option B (the sure payment) is different in every situation. In each row please select the option which most closely corresponds to your preference. As you go down the rows, if in some row you choose Option A (the draw) over Option B (the sure payment) please maintain this choice in the following rows (where the sure payment decreases). Note that the sums of money mentioned in this screen are hypothetical and your choice will not influence your final payoff.

			Option A	Option B	
	50% Chance	50% Chance			100% Chance
1)	£300	£0	<input type="radio"/>	<input type="radio"/>	£300
2)	£300	£0	<input type="radio"/>	<input type="radio"/>	£290
3)	£300	£0	<input type="radio"/>	<input type="radio"/>	£280
...					
29)	£300	£0	<input type="radio"/>	<input type="radio"/>	£20
30)	£300	£0	<input type="radio"/>	<input type="radio"/>	£10
31)	£300	£0	<input type="radio"/>	<input type="radio"/>	£0

(a) Quantitative item for eliciting risk preferences.

How do you see yourself: are you a person who is generally willing to take risks, or do you try to avoid taking risks?

completely unwilling to take risks 0 1 2 3 4 5 6 7 8 9 10 very willing to take risks

(b) Qualitative item for eliciting risk preferences.

A.2 Time preferences

Figures A2a and A2b depict the basic interface for eliciting time-discounting preferences - as implemented in Study 1. In Study 2, where we introduce an incentive compatible payment scheme, we stress that the monetary amounts are real and remove the hypothetical framing. Although we keep the numerical values of the outcomes the same, we remove the £symbol and specify that there is a 1 to 10 conversion rate to pounds so that 100 monetary units correspond to £10. In Study 3, we change the context from general to financial decision making. The text for the qualitative item for risk preferences in study 3 read as follows: *‘In comparison to others, are you a person who is willing to save money today in order to benefit from the financial gains of this investment in the future or are you not willing to do so ?’*

Figure A2: Interface for eliciting time-discounting preferences.

Suppose you were given the choice between the following: receiving a payment today or a payment in 12 months. We will present to you different situations - each row represents a different situation. Option A ("Payment today") is the same in all situations. Option B ("Payment in 12 months") is different in every situation. In each row please select the option which most closely corresponds to your preference. As you go down the rows, if in some row you choose Option B ("Payment in 12 months") over Option A (Payment today) please maintain this choice in the following rows (where the "Payment in 12 months" increases). Note that the sums of money mentioned in this screen are hypothetical and your choice will not influence your final payoff.

		Option A	Option B	
	Payment today			Payment in 12 months
1)	£100.0	<input type="radio"/>	<input type="radio"/>	£100.0
2)	£100.0	<input type="radio"/>	<input type="radio"/>	£103.0
3)	£100.0	<input type="radio"/>	<input type="radio"/>	£106.1
...				
23)	£100.0	<input type="radio"/>	<input type="radio"/>	£176.9
24)	£100.0	<input type="radio"/>	<input type="radio"/>	£180.9
25)	£100.0	<input type="radio"/>	<input type="radio"/>	£185.0

(a) Quantitative item for eliciting time-discounting preferences.

In comparison to others, are you a person who is generally willing to give up something today in order to benefit from that in the future or are you not willing to do so?

completely unwilling to give up something today very willing to give up something today

0 1 2 3 4 5 6 7 8 9 10

○

(b) Qualitative item for eliciting time-discounting preferences.

A.3 Altruism


Figures A3a and A3b depict the basic interface for eliciting preferences for altruism as implemented in Study 1. In Study 2, where we introduce an incentive compatible payment scheme, we stress that the monetary amounts are real and remove the hypothetical framing. We also adjust the numerical values to range from 0 to 100. We also remove the £symbol and specify that there is a 1 to 10 conversion rate to pounds so that 100 monetary units correspond to £10. In Study 3, we change the context from general to financial decision making. The text for the qualitative item for risk preferences in study 3 read as follows: ‘*How do you assess your willingness to share with others without expecting anything in return when it comes to donating money to charity?*’

Figure A3: Interface for eliciting risk preferences.

Imagine the following situation: you won £1,000 in a lottery.
Considering your current situation, how much would you donate to charity?
Note that the sums of money mentioned in this screen are hypothetical and your choice will not influence your final payoff.

0 50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800 850 900 950 1000


Select £ amount



(a) Quantitative item for eliciting preferences for altruism.

How do you assess your willingness to share with others without expecting anything in return when it comes to charity?

completely unwilling to share 0 1 2 3 4 5 6 7 8 9 10 very willing to share



(b) Qualitative item for eliciting preferences for altruism.