Order effects in eliciting preferences

Orestis Kopsacheilis [a,b,c] Sebastian J. Goerg [a]

Abstract

It has been recently suggested that combining responses from quantitative behavioral tasks and qualitative self-reports can enhance the accuracy of elicited preferences. Yet, how the order in which these items are encountered remains unexplored. Through three studies with 3,000 subjects, we test for order effects in preferences for risk, time-discounting, and altruism under two conditions: '*Quantitative First*' and '*Qualitative First*'. Results reveal systematic order effects: qualitative-first elicitations increase inferred patience and altruism, while quantitative-first enhances cross-method consistency for risk and time preferences. Monetary incentives mitigate time-discounting magnitude effects but not altruism. Additionally, framing qualitative items in financial contexts improves consistency.

[a]TUMCS for Biotechnology and Sustainability, Technical University of Munich, Straubing
[b]TUM School of Management, Technical University of Munich, Munich
[c]Corresponding author: 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. Accurately eliciting individuals' preferences along these dimensions carries significant implications across various applications, making it crucial for designing and developing better institutions and organizations. For example, knowledge about risk, time discounting, and altruistic preferences helps institutions design healthcare plans, pension schemes, and redistribution policies better suited to the citizens they set out to service. Similarly, organizations and management boards can utilize this information to develop products better suited to their customers, design more appropriate dynamic incentive schemes, and allocate their personnel into teams more efficiently.

We can broadly distinguish between two approaches in eliciting such preferences. The 'quantitative' approach infers preferences by observing people's choices, typically involving options expressed in monetary units. Conversely, the 'qualitative' approach involves asking people to self-assess and directly report their preference profile through some scale. An advantage often associated with quantitative items is their amenability to monetary incentivization, rendering them robust to measurement issues like inattention or the 'talk is cheap' criticism, wherein individuals respond based on social or self-image concerns rather than their actual preferences under monetary consequences. Conversely, qualitative items are easier to explain to subjects as well as faster and cheaper to implement. These qualities have contributed to the widespread use of such hypothetical survey items in large-scale surveys such as the German Socio-Economic Panel or the National Longitudinal Study of Youth (US-based).

Ideally, these two approaches would lead to similar conclusions. However, substantial evidence indicates otherwise, with several studies reporting notably low correlations between quantitative and qualitative measures—especially in the domain of risk preferences where this question has been

primarily investigated (e.g., Lönnqvist et al. 2015; Pedroni et al. 2017; Frey et al. 2017; Holzmeister and Stefan 2021).[1] This preference 'elicitation puzzle' raises important questions at both theoretical levels (regarding the nature of preferences and how to optimally model those) and practical levels (which approach is more predictive of future choices).

One possible explanation for this puzzle is that human preferences are multifaceted and, therefore, cannot be fully captured by the response to a single item. Indeed, recent evidence suggests that combining responses collected from a variety of different items can reduce measurement error and increase the correlation between experimental data and behavior observed in the field (Haesevoets et al., 2020, 2022; Wang and Navarro-Martinez, 2023).[2] A prominent example of such a multi-layered elicitation approach is the 'preference survey module' (Falk et al., 2023) and its streamlined variant that was used in the 'Global Preference Survey', which has already been utilized in one of the most extensive preference elicitation efforts, involving approximately 80,000 people across 76 countries (Falk et al., 2018). Seeking to harness the best of two worlds, this method elicits several preference dimensions (risk aversion, time discounting, trust, altruism, as well as positive and negative reciprocity) through both qualitative and quantitative items and then aggregates the two into a single score.

In this paper, we investigate whether the order in which qualitative and quantitative measures are presented influences the inference of preferences. Unlike previous studies, such as those by

[1] This literature has sometimes used the terms 'task' and 'ask' to distinguish between quantitative and qualitative items at the level of elicitation. Throughout this paper, we use the broader term 'item' to refer to both. Additionally, at the level of responses, the terms 'behavioral measures' and 'self-reports' are sometimes used to differentiate between quantitative and qualitative measures. Throughout this manuscript, whenever these terms are mentioned, they are always accompanied by the qualifiers 'quantitative' and 'qualitative'.

[2] Although several studies report significant correlations between experimentally elicited preference measures and real-world economic decisions—including occupational choice (Bonin et al., 2007), saving behavior (Sutter et al., 2013), loan repayment (Karlan, 2005), charitable donations (Benz and Meier, 2008), and health-related decisions such as physical activity (Chabris et al., 2008; Sutter et al., 2013)—these correlations tend to be weak. This weakness is particularly pronounced for risk (Weber et al., 2002; Sutter et al., 2013) and social preferences (Voors et al., 2012; Galizzi and Navarro-Martinez, 2019), likely due to their strong dependence on subjective perceptions and high domain specificity (e.g., Binswanger and Carman 2012; Isler et al. 2020; see also Levitt and List 2007 for a broader discussion).

Pedroni et al. (2017) and Falk et al. (2018), which employed these measures in a fixed order, our study systematically manipulates the order at the treatment level. Furthermore, extending the analysis of Pedroni et al. (2017), which focused exclusively on risk, we expand the scope to include time discounting and altruism.

Order effects occur when prior experience with one part of the experiment (or survey) affects responses in a subsequent part. Some of the most commonly suggested mechanisms underpinning them include 'anchoring' (Slovic 1967; Tversky and Kahneman 1974; see Furnham and Boo 2011, for a review), 'priming' (Bargh et al. 1996, but see also Weingarten et al. 2016 for a more recent meta-analysis), a 'preference for consistency' (Cialdini, 1984; Falk and Zimmermann, 2013, 2017), as well as several types of 'wealth effects' (Thaler and Johnson, 1990). To control for them, researchers often prefer between-subjects protocols, where subjects are exposed to one condition only, to avoid potential interaction effects, even at the cost of additional statistical power that would come from a within-subjects protocol (see Charness et al. 2012 for a discussion). More broadly, randomizing the order of items within a treatment and the order of treatments in within-subjects protocols is almost always considered good practice. Moreover, whenever subjects face multiple monetarily incentivized items, mechanisms such as the random incentive lottery are in place to control various types of wealth effects (Starmer and Sugden, 1991).

Despite efforts to prevent order effects from occurring, it is still important to investigate them in the context of eliciting preferences for at least two reasons: First, randomization may not always be possible, not least because of logistical challenges that come with large-scale experiments or surveys Two relevant examples of this are Pedroni et al. (2017) where quantitative items were always elicited first and Falk et al. (2018) where the opposite was the case. In such cases, knowing if and how order effects will likely affect conclusions about preferences would be informative. Second, discovering order effects and analyzing how they arise can yield valuable insights into the nature of preferences

with significant conceptual and practical implications.

Magnitude order effects occur when exposure to certain stimuli in one part of the experiment (or survey) affects the elicited preference by increasing or decreasing its reported intensity. For example, scaling up the monetary stakes of risky options in subsequent tasks has been shown to affect measurements of risk aversion beyond what can be attributed solely to risk preferences (Harrison et al., 2005; Holt and Laury, 2005).

Consistency order effects occur when the order in which two (or more) items are encountered affects the correlation between the measures associated with those tasks. For example, when subjects are asked to rate their life-satisfaction and their dating frequency over the past month, the answers to these questions exhibit differing degrees of correlation depending on the order they were asked (Strack et al., 1988; Kahneman et al., 2006).

We conduct online experiments in three studies with a total of 3,000 subjects, where we elicit preferences about risk, time discounting and altruism in variations of two conditions. In *'Quantitative First'*, quantitative items are elicited before qualitative ones for every preference dimension, while in *'Qualitative First'*, the opposite is the case. Both qualitative and quantitative items are based on adaptations of the preference survey module (Falk et al., 2023). This module elicits quantitative measures for risk and time-discounting through hypothetically incentivized multiple price lists and uses a hypothetical donation question for altruism. Qualitative measures are always elicited through a simple self-assessment question.

Across three studies we investigate how manipulations related to the incentivization and framing of these preferences can account for magnitude and consistency order effects. Adding to previous literature reporting low cross-method correlation in the domain of risk, our findings reveal that preferences elicited from quantitative and qualitative methods also exhibit relatively low correlations in the domains of altruism and time-discounting. However, the more relevant contribution of

this paper is that we observe systematic order effects. Concerning magnitude order effects, eliciting preferences through qualitative items first enhances inferred patience and altruism. The incentive-compatible implementation of quantitative measures inoculates time preferences from such effects but not preferences regarding altruism. Regarding consistency order effects, using quantitative items first increases the cross-method correlation for risk and time preferences. Interestingly, re-framing qualitative items in the context of financial decision-making, such asymmetries disappear for time discounting preferences and are mitigated for risk preferences. We discuss the implications of our findings in the context of nudging interventions aiming to increase charitable giving as well as for our understanding of the nature of preferences.

2 Methods

We conduct three studies, each with 1,000 subjects that we recruit online through Prolific Academic. The median completion time for sessions in Studies 1 and 3 was four minutes while for Study 2, it is five minutes. Every subject received a flat fee of £1 for their participation in Study 1 and £0.8 in Studies 2 and 3. Additionally, in Study 2, one out of twenty subjects had one of their answers in the quantitative questions (across all three preference dimensions) played out for real, which earned those subjects an additional £16 - on average. In Study 2, we implement a 1 to 10 exchange rate between the monetary units presented on screen and the actual payoff. The monetary units on screen correspond to the same values as in Studies 1 and 3 with the exception of the preference for altruism, where we compress the range to 1-100 instead of 1-1000.

In every study, we randomly assign subjects to one of two treatments: *'Quantitative First'* or *'Qualitative First'*. The only difference between the two treatments is that in the *Quantitative First* (*Qualitative First*) treatment, subjects encounter the quantitative (qualitative) item before the qualitative (quantitative) one.

In Study 1 we implement the baseline preference survey module, in Study 2 we introduce incentive-compatibility for quantitative items, while in Study 3 we introduce financial context when formulating the questions in qualitative items. In each study, subjects encounter the preference dimension in a randomized order. However, within each preference dimension, the order in which the qualitative and quantitative items are encountered is fixed and dependent on the treatment.

The items we use across these three studies are based on adaptations from the preference survey module, as described in Falk et al. (2023). This module elicits preferences on risk, time-discounting, altruism, and trust, as well as positive and negative reciprocity through a series of quantitative and qualitative survey items. Both quantitative and qualitative items are hypothetically incentivized. These items have been selected—among a broader battery of survey items that were also elicited—based on their ability to predict choices in monetarily incentivized experiments that were conducted with the same subjects in sessions one week apart.

Qualitative items consist of questions asking subjects to self-assess their preference (either in general or compared to others) while quantitative items are hypothetical versions of the task included in the incentivized choice experiment. Except for positive reciprocity, the module is symmetric, containing a quantitative and a qualitative item for each preference dimension.

Table 1 provides an overview of the preference survey module and its adaptations across our three studies. [3] We focus on three preference dimensions: risk, time-discounting, and altruism. 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. Importantly, in our implementation of the preference survey module, we vary the order in which quantitative and qualitative items are encountered on a treatment level. Falk et al. (2023), when validating the preference survey module, did not randomize this order. There, qualitative items were always

[3] For more details regarding the interface of each item, see Appendix A.

elicited first for all preference dimensions, and only then quantitative items are administered. That same, fixed order is maintained in the implementation of this module in the Global Preference Survey (Falk et al., 2018).

Table 1: The preference survey module adapted from (Falk et al., 2023)

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 'today' 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'. Detailed instructions can be found in the Appendix.

The quantitative item for eliciting risk preferences consists of a multiple-price list. Subjects make 31 choices between a lottery offering a high and a low outcome with equal (50%) chance and a safe option offering a monetary amount with certainty (100% chance). The lottery remains constant across choices while the safe option ranges from the highest (£300) to the lowest amount (£0) offered by the lottery. Study 2 introduces a 1 to 10 exchange rate according to which the high outcome is worth £30 while the safe option ranges from £30 to £0. Although we keep the numerical values of the outcomes that subjects see on-screen the same across the three studies, we remove the '£' symbol from Study 2 and specify that there is a 1 to 10 conversion rate between

the monetary units seen on screen and their actual potential payoff. This specification is similar for the quantitative items of time and altruism too. This type of task is commonly used in the literature to infer risk preferences (see Andersen et al. 2006 for a review). Here, as well as in the multiple price list for time-preferences, we impose a single-switching point and, therefore, prevent subjects from committing violations of monotonicity that are commonly attributed to errors from inattention. 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. There 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 Starmer (2000) for a more extensive overview.

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 while that in 12 months ranges from £100 to £185. According to the 1 to 10 exchange rate we introduce in Study 2, the payment today is worth £10 while that in 12 months ranges from £10 to £18.5. Subjects are informed that if the answer that is randomly selected for payment indicates a preference for receiving payment after 12 months, then they will only receive their bonus payment at that later time. Switching from the immediate payment to that in 12 months '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. Just as in risk preferences, however, there is a wealth of models incorporating more nuanced behavioral insights related to time preferences (Laibson, 1997; Frederick et al., 2002).

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 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 (£1000) to oneself while the other end is to charity. In Study 2, we re-scale the initial endowment that subjects see down to 100 monetary units. Given the 1 to 10 conversation rate that we implement in Study 2, this now corresponds to a value of £10.

In the experimental literature, altruistic behavior has most commonly been studied using the dictator game, wherein a player decides how much of an endowment to retain for themselves and how much to transfer to a second, passive party (Forsythe et al. 1994; see also List 2007; Bardsley 2008 for a critical discussion on the interpretation of giving in dictator games). However, variations like the one implemented here, in which the second party is replaced with a non-profit institution outside the laboratory, are also common (Eckel and Grossman, 1996). Irrespective of the task employed to elicit charitable giving, the consensus is that individuals exhibit a preference for giving—a conclusion that challenges the neoclassical account of a selfish agent concerned solely with maximizing individual payoff.

For the qualitative items we always use scales that allow for different levels of agreement to a certain statement. The statements can be seen in 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 qualitative self-reports are based on items in existing surveys, like the German Socio-Economic Panel Study, the National Longitudinal Study

of Youth as well as previous research (e.g., Weber et al. 2002; Perugini et al. 2003).

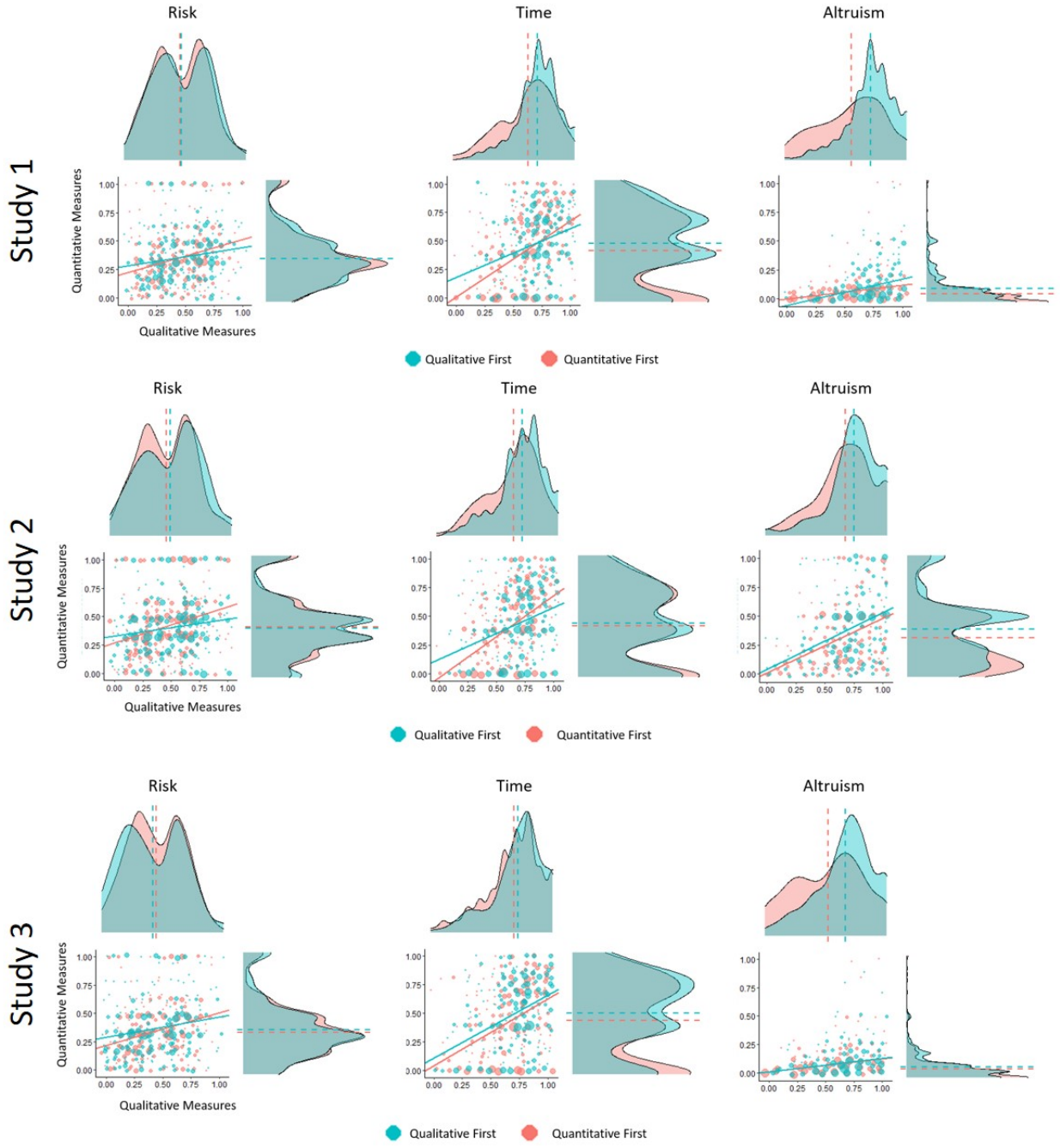
The final measures for all six items are standardized between 0 and 1. For the qualitative measures, we take the degree of agreement and divide it by the range of the scale ('10' for all three survey items). Similarly, for the quantitative measure of altruism, we divide the amount contributed by the 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 by 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 in 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 pre-analysis plan and recruitment protocol is common for all 3 studies and has been pre-registered. The pre-registration can be accessed at https://aspredicted.org/44K_1R5.

3 Results

Figure 1 provides a visual impression of our data across preference dimensions, treatments, and studies. The y-axis captures the measure of an individual in the quantitative measure while the x-axis represents that for the qualitative measure. Points with light-blue fill correspond to measurements in the *Qualitative First* treatment while those with light red derive from the *Quantitative*

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.

First treatment. The constant and slope of the plotted linear models derive from Ordinary Least Squares regressions of quantitative on qualitative measures.

Focusing on Study 1 (top row), we can see from the partial density plots that distributions of preference measures within each elicitation method are 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 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 patterns 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, disappears in Study 3. This suggests that (at least part of) the apparent dissonance between the two elicitation methods is due to differences in context. Our statistical analysis supports these visual impressions.

3.1 Magnitude order effects

Table 2 focuses on magnitude order effects, reporting the averages for each measure, as well as the p-values derived from the Mann-Whitney (MW) tests. In Study 1, where quantitative items

are hypothetically incentivized and qualitative 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: $p = 0.001$; altruism: $p < 0.001$). There is no significant difference for risk ($p = 0.592$). The same tendency is observed for measures obtained from qualitative items. Subjects in *Qualitative First* self-report higher degrees of patience ($p < 0.001$) and altruism ($p < 0.001$) compared to *Quantitative First*. Again, risk preferences are impervious to such magnitude order effects ($p = 0.249$).

Table 2: 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. Average measures for elicited preferences. 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 ($p = 0.549$). In addition, we see

that monetary incentivization in an incentive-compatible way inoculates quantitative measures from magnitude order effects in time discounting preferences ($p = 0.274$), which were present in Study 1. However, people are still behaving more altruistic in *Qualitative First* compared to *Quantitative First* ($p < 0.001$). 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 *Qualitative First*. These effects are consistently statistically significant at 5% level in MW tests (risk: $p = 0.016$; time: $p < 0.001$; altruism: $p < 0.001$).

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, the magnitude order effect reappears in time preferences for quantitative measures ($p = 0.004$). 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 ($p = 0.577$) while there are statistically significant ones for altruism ($p < 0.001$). 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: $p = 0.014$; altruism: $p < 0.001$). Risk preferences are an exception, 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: $p = 0.057$).

Our nominal comparisons and statistical analyses are conducted across treatments and within

each study. Nonetheless, it is worth noting that average measures are remarkably consistent across studies for each preference dimension and each elicitation method. One exception to this concerns 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 most likely because the scale was readjusted from 1-1000 (Study 1 and Study 3) to 1-100 (Study 2).

3.2 Consistency order effects

Table 3 summarises the Pearson and Spearman-rank correlations between quantitative and qualitative measures. To test whether these correlations were statistically significant across treatments, we conduct Fisher z-tests and report their p-values. 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 ($p = 0.070$) and patience ($p < 0.01$) but not for altruism ($p = 0.295$).

This pattern remains the same in Study 2, with the introduction of monetary incentives (risk: $p = 0.028$; time: $p < 0.001$; altruism: $p = 0.467$). However, 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* ($p = 0.083$), the cross-method difference in correlation for time preferences is no longer statistically significant ($p = 0.607$). Like in Studies 1 and 2, the cross-method correlation for altruism does not differ significantly across treatments ($p = 0.607$).

Table 3: Consistency order effects

Study 1	*Quantitative measures: Hypothetical incentives/ Qualitative measures: General context*				
	Quant-First		Qual-First		Fisher z-test
	Pearson	Spearman	Pearson	Spearman	(p-value)
Risk	0.274	0.297	0.165	0.228	0.070
Time	0.485	0.483	0.268	0.274	0.000
Altruism	0.284	0.334	0.344	0.365	0.295
Study 2	*Quantitative measures: Incentive compatible/ Qualitative measures: General context*				
Risk	0.276	0.296	0.144	0.189	0.028
Time	0.486	0.488	0.282	0.280	0.000
Altruism	0.405	0.400	0.366	0.367	0.467
Study 3	*Quantitative measures: Hypothetical incentives/ Qualitative measures: Financial context*				
Risk	0.285	0.317	0.181	0.216	0.0831
Time	0.409	0.409	0.382	0.344	0.607
Altruism	0.306	0.410	0.264	0.329	0.471

Note. Pearson and Spearman correlations between quantitative and qualitative measures. '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. P-values are obtained from Fisher z-tests testing for differences between the 'Quant-First' and 'Qual-First' treatments.

4 Discussion

Eliciting preferences accurately is of vital importance for social scientists and practitioners alike. Concerningly, different methods of eliciting such preferences often lead to different conclusions (e.g. Pedroni et al. 2017). Moreover, the correlation between experimental or survey measures and economic behavior observed in the field is not always strong (Sutter et al., 2013; Galizzi and Navarro-Martinez, 2019). To remedy these problems, researchers increasingly employ methods that elicit preferences through a diverse set of items. A prominent example of such a method is the 'preference survey module' (Falk et al., 2023) that combine qualitative (self-assessment questions) and quantitative (choices between options framed in financial terms). A previously unaddressed question with important methodological and theoretical consequences is whether the order in which people respond to these items influences elicited preferences.

In this paper, we test for the presence of order effects when preferences are elicited using quantitative behavioral tasks and qualitative self-reports within the same session. Previous studies that have employed both types of items have typically done so in a fixed order with either qualitative items preceding quantitative ones (e.g. Falk et al. 2023) or the other way around (e.g. Pedroni et al. 2017). In our study instead, we manipulate this order between treatments. In the *Quantitative First* treatment, quantitative items precede qualitative ones while in *Qualitative First*, the opposite is the case.

Across three studies, we find consistent evidence suggesting that order effects are statistically significant and substantial, affecting both the level of the elicited measures (magnitude) as well as their cross-method correlation (consistency).

Concerning magnitude order effects, we find that when incentives are hypothetical (Studies 1 and 3), responding in qualitative self-assessments first increases inferred patience and altruism in quantitative and qualitative measures, compared to when people go through the quantitative item first. Either anchoring or a preference for consistency plausibly explain this pattern. Even though the two scales are incompatible, responses were systematically closer to the upper end of the scale in qualitative rather than quantitative items. Therefore, for instance, a subject self-reporting a score near the maximum of the available scale for altruism might be anchored or feel compelled by an implicit sense of consistency to inflate their ensuing charitable donation compared to the alternative where the charitable donation precedes.

Moreover, in line with the 'talk is cheap' criticism often ascribed to qualitative measures, we observe that incentive compatibility (Study 2) inoculates quantitative (but not qualitative) time-discounting measures from magnitude order effects. Not all talk is cheap, however, as the quantitative measure for altruism remains susceptible to magnitude order effects. Specifically, we observe a 24% increase in charitable donations when individuals report their level of altruism before being

asked to donate compared to the reverse order. This result closely mirrors findings related to moral nudges. Capraro et al. (2019) found that when individuals are asked what they think is the morally right thing to do, subsequent charitable giving increases by 44% compared to when they are asked to contribute without this moral assessment. Similarly, Andreoni and Serra-Garcia (2021) demonstrate how individuals with time-inconsistent preferences can be prompted to increase one-time donations by up to 50% by linking this behavior to implicit norm conformity tied to social image concerns. We argue that the introspective nature of the self-reported qualitative items evokes normative considerations regarding the 'right' thing to do. In this sense, qualitative questions tap into a mechanism similar to that described in moral nudges. Exploring this mechanism further holds promise for future research with implications for charitable giving. For instance, in our study, subjects were allowed to contribute up to £10. Investigating how the scale of the donated amount influences these results in a follow-up study would be interesting.

To examine consistency order effects, we analyze cross-method correlations between quantitative and qualitative measures across treatments. At face value, the overall small to medium correlations we observe in Studies 1 and 2 echo the concerns raised by prior literature, pointing to fundamental incongruities between the two elicitation traditions. Interestingly, we observe that eliciting quantitative measures before qualitative ones significantly increases the cross-method correlation for risk and time preferences across Studies 1 and 2—but not altruism.

We suggest that this type of consistency order effect can be understood by noting that both studies present qualitative questions in a broad context (e.g., 'Are you a person who is **generally** willing to take risks?'). However, in Study 3, where qualitative measures are framed within a financial context (e.g., 'How do you evaluate your attitude towards risk regarding **financial investments**?'), the cross-method correlation in 'Qualitative First' improves. This change eliminates the asymmetry with 'Quantitative First' as the consistency order effects observed in Studies 1 and 2

are no longer statistically significant in Study 3. A similar tendency is observed for risk-preferences, but the consistency order effect is still significant in Study 3, albeit only at the 10% level.

These results contribute to the growing body of evidence indicating that judgments and preferences are influenced by the context in which they were elicited (Tversky and Thaler, 1990; Tversky and Simonson, 1993; Loewenstein, 1999; Lichtenstein and Slovic, 2006; Barseghyan et al., 2011; Dohmen et al., 2011; Stewart et al., 2015). They thus corroborate our intuition that (at least part of) the reason for this dissonance is the general (or abstract) context in which qualitative items are typically presented in surveys. To illustrate this intuition, consider the qualitative risk-preference item, which asks respondents how risk-tolerant they are 'in general'. This question can prompt scenarios of financial decision-making for some and health decisions for 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. Conversely, when respondents are primed with the financial context embedded in quantitative questions, that context carries over to the general framing of the qualitative item, enhancing the cross-method correlation. This echoes the finding that when a specific question precedes a general one, respondents use the information primed by the specific question to form the general judgment (Schwarz et al., 1991). On the other hand, it is worth noting that the cross-method correlation for altruism remained unaffected across all three studies, adding to the evidence that behavior in dictator games is insensitive to framing manipulations (see also Dreber et al. 2013; Goerg et al. 2020).

The consistency order effects observed for risk and time preferences have two important implications. First, assessing the consistency between these two elicitation traditions using a general frame in qualitative items—as is often the case—likely overstates the incongruity between the measures.

Combined with 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'. In light of these results, preferences are likely more consistent than some recent findings have suggested. Second, our finding that quantitative items involving monetary trade-offs highlight a financial decision-making context has important implications for improving survey design. Given that preferences often exhibit significant context dependencies, our main recommendation is that when surveyors' primary interest lies in non-financial domains (e.g., health decisions), they should explicitly frame questions in terms of those specific contexts.

Declarations

Conflict of interest The authors have no conflict of interest.

Ethics approval This study was approved by the German Association for Experimental Economic Research e.V. Institutional Review Board Certificate No. 2uUA3J8W.

https://gfew.de/ethik/2uUA3J8W

References

Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2006). Elicitation using multiple price list formats. *Experimental Economics*, 9:383–405.

Andreoni, J. and Serra-Garcia, M. (2021). Time inconsistent charitable giving. *Journal of Public Economics*, 198:104391.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental economics*, 11:122–133.

Bargh, J. A., Chen, M., and Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of personality and social psychology*, 71(2):230.

Barseghyan, L., Prince, J., and Teitelbaum, J. C. (2011). Are risk preferences stable across contexts? evidence from insurance data. *American Economic Review*, 101(2):591–631.

Benz, M. and Meier, S. (2008). Do people behave in experiments as in the field?—evidence from donations. *Experimental economics*, 11:268–281.

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.

Binswanger, J. and Carman, K. G. (2012). How real people make long-term decisions: The case of retirement preparation. *Journal of Economic Behavior & Organization*, 81(1):39–60.

Bonin, H., Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2007). Cross-sectional earnings risk and occupational sorting: The role of risk attitudes. *Labour Economics*, 14(6):926–937.

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.

Chabris, C. F., Laibson, D., Morris, C. L., Schuldt, J. P., and Taubinsky, D. (2008). Individual laboratory-measured discount rates predict field behavior. *Journal of risk and uncertainty*, 37:237–269.

Charness, G., Gneezy, U., and Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization*, 81(1):1–8.

Cialdini, R. B. (1984). *Influence: The new psychology of modern persuasion*. Morrow.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the european economic association*, 9(3):522–550.

Dreber, A., Ellingsen, T., Johannesson, M., and Rand, D. G. (2013). Do people care about social context? framing effects in dictator games. *Experimental Economics*, 16:349–371.

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.

Falk, A. and Zimmermann, F. (2013). A taste for consistency and survey response behavior. *CESifo Economic Studies*, 59(1):181–193.

Falk, A. and Zimmermann, F. (2017). Consistency as a signal of skills. *Management Science*, 63(7):2197–2210.

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.

Furnham, A. and Boo, H. C. (2011). A literature review of the anchoring effect. *The journal of socio-economics*, 40(1):35–42.

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.

Goerg, S. J., Rand, D., and Walkowitz, G. (2020). Framing effects in the prisoner's dilemma but not in the dictator game. *Journal of the Economic Science Association*, 6:1–12.

Haesevoets, T., Folmer, C. R., and Van Hiel, A. (2022). Improving the measurement of prosociality through aggregation of game behavior. *Current Opinion in Psychology*, 44:237–244.

Haesevoets, T., Van Hiel, A., Dierckx, K., and Folmer, C. R. (2020). Do multiple-trial games better reflect prosocial behavior than single-trial games? *Judgment and Decision Making*, 15(3):330–345.

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. (2005). Risk aversion and incentive effects: New data without order effects. *American Economic Review*, 95(3):902–912.

Holzmeister, F. and Stefan, M. (2021). The risk elicitation puzzle revisited: Across-methods (in) consistency? *Experimental Economics*, 24(2):593–616.

Isler, O., Isler, B., Kopsacheilis, O., and Ferguson, E. (2020). Limits of the social-benefit motive among high-risk patients: a field experiment on influenza vaccination behaviour. *BMC Public Health*, 20(1):1–9.

Kahneman, D., Krueger, A. B., Schkade, D., Schwarz, N., and Stone, A. A. (2006). Would you be happier if you were richer? a focusing illusion. *science*, 312(5782):1908–1910.

Karlan, D. S. (2005). Using experimental economics to measure social capital and predict financial decisions. *American Economic Review*, 95(5):1688–1699.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478.

Levitt, S. D. and List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives*, 21(2):153–174.

Lichtenstein, S. and Slovic, P. (2006). The construction of preference: An overview. *The construction of preference*, 1.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political economy*, 115(3):482–493.

Loewenstein, G. (1999). Experimental economics from the vantage-point of behavioural economics. *The Economic Journal*, 109(453):25–34.

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.

Schwarz, N., Strack, F., and Mai, H.-P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public opinion quarterly*, 55(1):3–23.

Slovic, P. (1967). The relative influence of probabilities and payoffs upon perceived risk of a gamble. *Psychonomic Science*, 9(4):223–224.

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.

Starmer, C. and Sugden, R. (1991). Does the random-lottery incentive system elicit true preferences? an experimental investigation. *The American Economic Review*, 81(4):971–978.

Stewart, N., Reimers, S., and Harris, A. J. (2015). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science*, 61(3):687–705.

Strack, F., Martin, L. L., and Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European journal of social psychology*, 18(5):429–442.

Sutter, M., Kocher, M. G., Glätzle-Rützler, D., and Trautmann, S. T. (2013). Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *American Economic Review*, 103(1):510–531.

Thaler, R. H. and Johnson, E. J. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management science*, 36(6):643–660.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5:297–323.

Tversky, A. and Simonson, I. (1993). Context-dependent preferences. *Management science*, 39(10):1179–1189.

Tversky, A. and Thaler, R. H. (1990). Anomalies: preference reversals. *Journal of Economic Perspectives*, 4(2):201–211.

Voors, M., Turley, T., Kontoleon, A., Bulte, E., and List, J. A. (2012). Exploring whether behavior in context-free experiments is predictive of behavior in the field: Evidence from lab and field experiments in rural sierra leone. *Economics Letters*, 114(3):308–311.

Wang, X. and Navarro-Martinez, D. (2023). Increasing the external validity of social preference games by reducing measurement error. *Games and Economic Behavior*, 141:261–285.

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.

Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., and Albarracín, D. (2016). From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological bulletin*, 142(5):472.

A Instructions

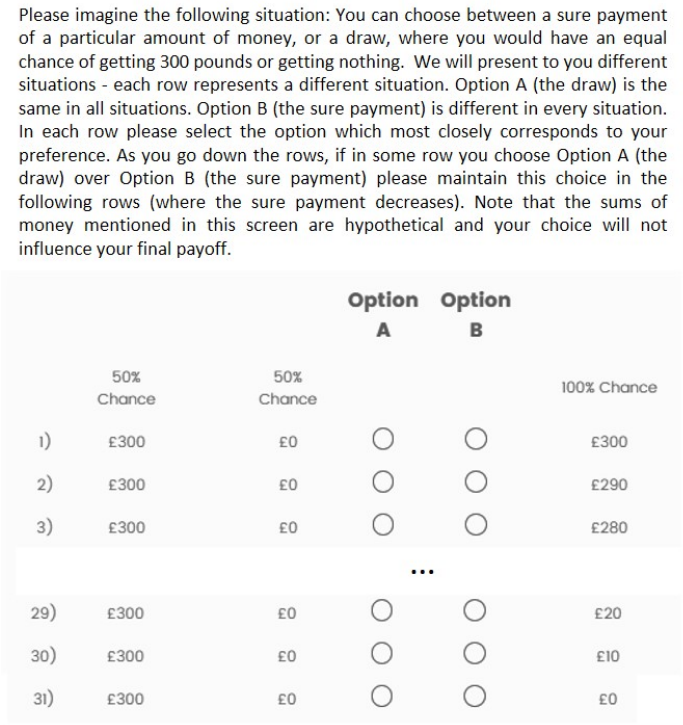
A.1 Risk preferences

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?*'

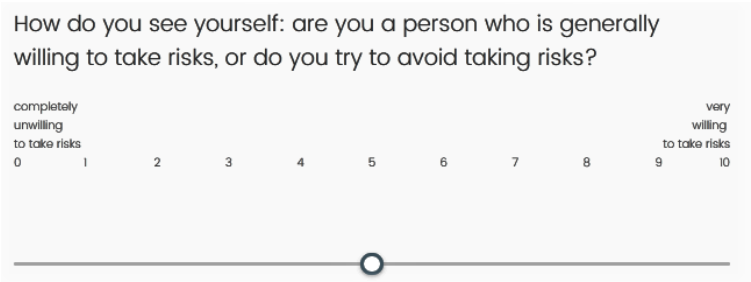
A.2 Time preferences

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 A1: Interface for eliciting risk preferences.



(a) Quantitative item for eliciting risk preferences.

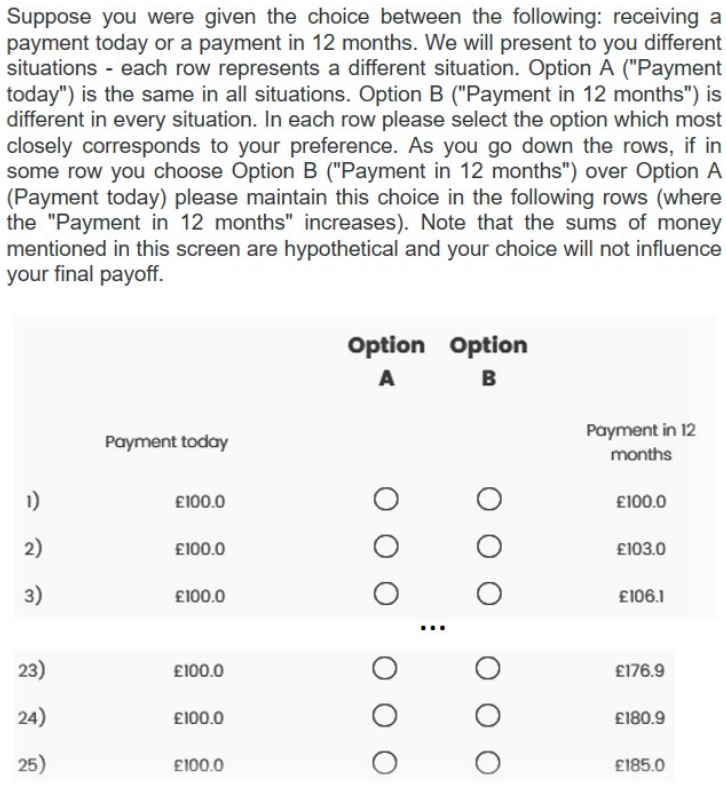


(b) Qualitative item for eliciting risk preferences.

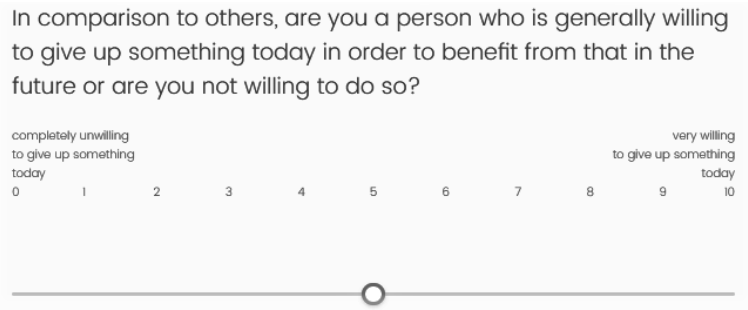
A.3 Altruism

Figures A3a and A3c depict the interface of the baseline preference survey module 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. Further, we adjust the numerical values to range from 0 to 100. We also remove the '£' symbol and specify a 1 to 10 conversion rate to pounds so that 100 monetary units cor-

Figure A2: Interface for eliciting time-discounting preferences.



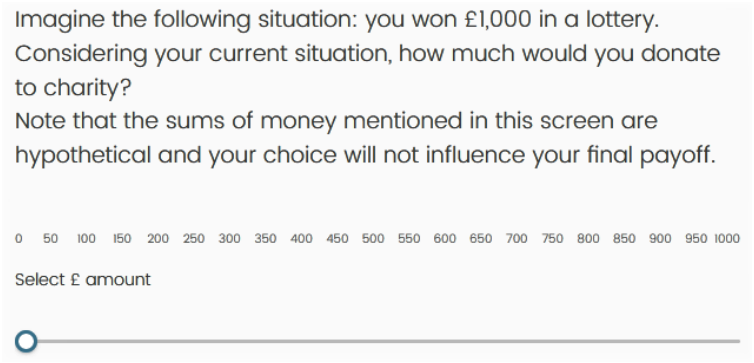
(a) Quantitative item for eliciting time-discounting preferences.



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

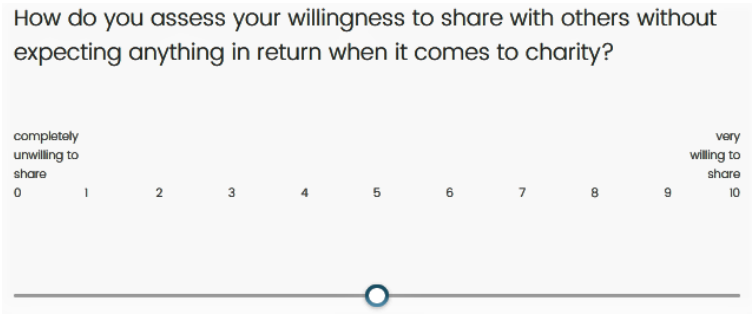
respond to £10. Lastly, we provide a drop-down menu with a list of well-known charities (Figure A3b). In Study 3, we change the context from general to financial decision-making. The text for the qualitative item for risk preferences in study 3 read as follows: '*How do you assess your willingness to share with others without expecting anything in return when it comes to donating money to charity?*'

Figure A3: Interface for eliciting risk preferences.



(a) Quantitative item for eliciting preferences for altruism.

(b) List of charities.



(c) Qualitative item for eliciting preferences for altruism.