Testing the Elicitation Procedure of the Minimum Acceptable Probability

Maria Polipciuc* Martin Strobel*

March 7, 2022

Abstract

*Maastricht University. Email: m.polipciuc@maastrichtuniversity.nl. We thank Elias Tsakas and participants in the BEELab proposal meeting for valuable comments.

1 Introduction

Individuals have often been found to prefer exposure to a randomly generated risk than to an equiprobable risk generated by an opponent in a strategic situation. In the context of trust games, this strategic risk premium has been dubbed *betrayal aversion* (Bohnet and Zeckhauser, 2004). Many papers find that betrayal aversion is an important determinant of trust CITE.

Betrayal aversion is identified as the difference in first mover behavior in two games: a binary trust game CITATION and an equivalent game where the decision at the second node is made by a random device. First movers typically have to indicate their *minimum acceptable probability* (MAP). This is elicited using a strategy-method procedure CITE and has many features in common with the Becker–DeGroot–Marschak mechanism (BDM). The MAP is a first mover's conditional threshold probability of the good outcome at the second node for preferring to send money to the second mover over the outside option (which keeps the two players' equal initial endowments unchanged). It is elicited without first movers knowing how many second movers (devices) chose the favorable outcome at the second node. Then, should the threshold be reached or exceeded, first movers send money to second movers, and the outcome is decided by their matched second mover's choice (their matched device's choice). Should the threshold not be met, the outside option is implemented.

A recent paper has shown theoretically that the elicitation procedure of MAPs used in most papers on betrayal aversion leaves the door open to potential confounds such as "ambiguity attitudes, complexity, different beliefs, and dynamic optimization" if players are not rational expected utility maximizers (Li et al., 2020). Moreover,

a couple of empirical papers which use more stringent identification procedures for betrayal aversion by controlling for beliefs in the two games do not find betrayal aversion (Fetchenhauer and Dunning, 2012, the second experiment in Polipciuc, 2021), or find it to play a role for trusting only when beliefs are far more optimistic than is generally the case (Engelmann et al., 2021). Several papers CITE find that when dealing with complex risks, participants in experiments require an extra premium compared to simple risk aversion. This premium is positively correlated with ambiguity aversion. (ARE EFFECT size similar?)

In this note, we use an online experiment to measure how much of what has been called betrayal aversion is due to distributional dependence, regardless of the source of risk being random or strategic. We remove the strategic component and show participants complete distributions over probabilities of the good (and bad) outcome of a lottery, and ask them to state their cutoff probability of the favorable outcome for preferring the lottery over a safe payoff. Since this is a situation involving complex risk, we expect a premium between the distribution mimicking the control condition in betrayal aversion studies and the distribution mimicking the binary trust game condition, as suggested by Li et al. (2020). We find the opposite to be true: the higher the expected value of the probability of the favorable outcome, the higher the minimum acceptable probability required by participants to accept the lottery.

While this is at odds with our expectations, it ties in with findings from the empirical literature on distributional dependence of willingness to pay (WTP) as elicited through the BDM mechanism. Similarly to betrayal aversion, theoretical literature has pointed out that the BDM mechanism is not incentive compatible if players are not rational expected utility maximizers (Karni and Safra, 1987;

Horowitz, 2006). This is because individuals face uncertainty regarding the price of the good and additional uncertainty about whether they will buy the good or not. If their utility function is influenced by these uncertainties, changing the price distribution of the good might influence their MAP.

Several empirical papers find this to be the case for the BDM: generally, the higher the expected price of the good, the higher the WTP (for a short review of this literature, see Tymula et al., 2016). Here I plan to say something about possible things which may be causing the result. For this, I have to understand if two papers cited by Tymula et al. (2016)—Kőszegi and Rabin (2006) and Wenner (2015)—are relevant for our study, and what they imply for our findings.

The paper is structured as follows. Section 2 describes the experimental design and procedures. Section 3 sets forth the hypothesis. Section 4 presents the results. Section 5 explains how our results inform the existing literature.

2 Design and procedures

We use a within-subject design, with each subject being exposed to all treatments sequentially. In each treatment, participants see a distribution over a lottery with two possible outcomes: a high payoff and a low payoff. A lottery will be drawn at random from the distribution. This means in some treatments it is more likely to get a lottery with a high chance of a high payoff than in others. We use three distributions over lotteries. The distributions are ordered in terms of the expected payoff over the entire distribution, as their name suggests: the Good, the Bad, and the Uniform (the Good $>$ the Uniform $>$ the Bad).

Two of the three distributions are meant to emulate treatments in papers on

betrayal aversion. The Uniform distribution has equal chances of occurrence for each of the possible lotteries. We assume that this is what participants expect to face in treatments with decisions made by randomization devices, unless specified otherwise. The Bad distribution has an overall chance of a high payoff similar to the share of trustworthy respondents in papers on betrayal aversion (0.2895). The distribution in the Good treatment mirrors the one in the Bad treatment: its overall expected chance of a high payoff is one minus that in the Bad treatment (0.7105), it has the same variance and minus the skewness of the Bad distribution. We included this distribution to check if departures from the Uniform distribution in either direction yield effects of similar size (albeit reverse sign).

include table describing distributions

Table 1: The treatments: the distribution of chances (X in 15) of a high payoff

X	The Good	The Bad	The Uniform
0	1	8	2
1	1	4	2
2	1	4	2
3	1	3	2
4	1	2	2
5	1	1	2
6	1	1	2
7	1	1	2
8	1	1	2
9	1	1	2
10	1	1	2
11	2	1	2
12	3	1	2
13	4	1	2
14	4	1	2
15	8	1	2
Total	32	32	32

To make the task easy to understand, we present lotteries via 32 wheels of fortune with 15 sectors each. Dark blue sectors symbolize the high payoff (£4), light blue sectors—the low payoff (£1). The sure payoff participants received if no wheel is spun is £2. In each treatment, participants see the entire distribution of lotteries in that treatment, sorted in ascending order by the probability of the favorable outcome. Figure 1 below shows the distribution for the Good treatment.

Participants are told that one of the wheels will be drawn at random, with all wheels having an equal chance to be drawn. They are asked to state a *minimum acceptable frequency* (which we refer to as MAP, even though it is not a probability, but a frequency, for easier comparison with papers on betrayal aversion): the lowest number of dark blue sectors in the randomly drawn wheel such that they prefer to spin the wheel for their payoff instead of receiving the sure payoff.[1] Specifically, they have to answer: "Which wheels would you like to spin for your bonus?" by inserting an integer between 0 and 15 in the blank space: "I prefer to spin wheels which have at least _____ dark blue sectors."

The experiment was conducted online using Qualtrics. Participants were UK residents registered on a platform for conducting academic studies (Prolific). Since the elicitation of MAPs is rather complex (Quercia, 2016; Polipciuc and Strobel, 2020), we opted for participants who had at least a bachelor's degree. The study was pre-registered at the AEA RCT Registry (https://doi.org/10.1257/rct.7776-1.1).

Table 2 in Appendix ... describes the sample. Treatment was assigned in order to balance the number of participants exposed to each of the six possible orderings

[1]We decided to use frequencies instead of probabilities because there is evidence that participants have an easier time expressing choices this way (Quercia, 2016).

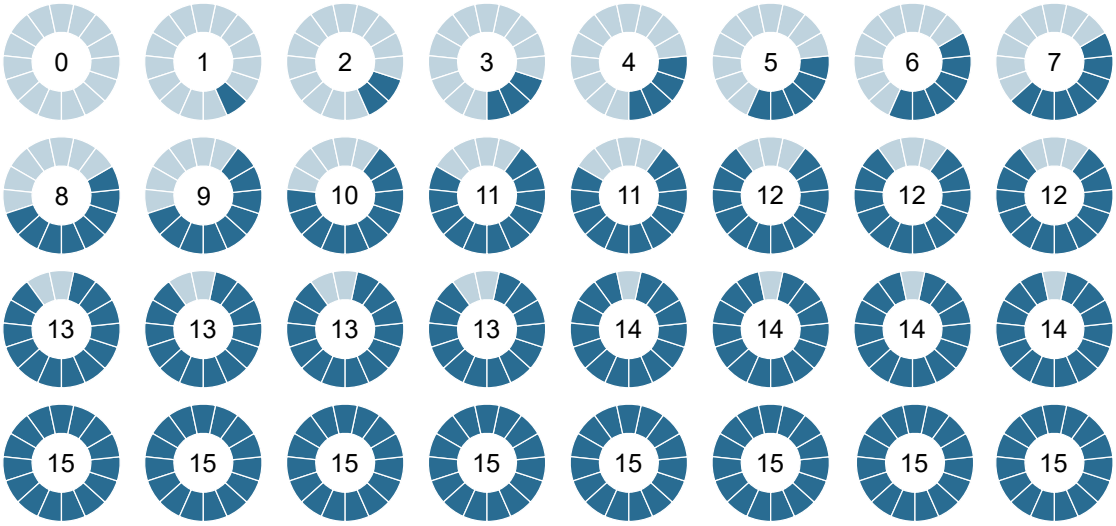

Figure 1: The Good distribution

of treatments. 275 of the 450 participants answered the eliminatory comprehension questions correctly and completed the experiment. Since assignment to treatment happened before participants had gone through the comprehension questions, this leads to slightly different sizes of the subsamples for the six orderings. The study had three stages: the eliminatory comprehension questions, the three decisions, and a post-experimental questionnaire.[2] Those who completed the experiment (did not complete the experiment) spent a median time of 12.4 (5.9) minutes and earned on average 3.96 (1) UK pounds.[3]

[2]In the post-experimental questionnaire, respondents answered an unincentivized question to determine their ambiguity aversion, an adapted cognitive reflection test (Frederick, 2005; Thomson and Oppenheimer, 2016), a question about the subject they studied for their most recent degree, a general risk taking question (Dohmen et al., 2011), a question about their aspiration level for earnings from participating in a survey, a couple of questions to check their anchoring susceptibility, from which an anchoring score can be computed (Cheek and Norem, 2017), a set of questions about their optimism/pessimism, the revised Life Orientation Test (Scheier et al., 1994) and a brief sensation seeking scale, BSSS-4 (Stephenson et al., 2003).

[3]Participants were paid £1 for going through the comprehension questions (regardless of the correctness of their answers). Those who answered correctly then earned an additional £1, £2 or £4 for one of their decisions.

The high average earnings of those who completed the experiment are due to a coding error. Instead of decisions in all three treatments being equally likely to be selected, only the Good and

Table 2: Characteristics of the estimation sample

	Age	Share male	Sample size
Good–Uniform–Bad	30.956	0.333	45
	(8.808)	(0.477)	
Uniform–Bad–Good	33.538	0.346	52
	(9.074)	(0.480)	
Bad–Good–Uniform	37.114	0.523	44
	(11.071)	(0.505)	
Good–Bad–Uniform	33.132	0.491	53
	(9.174)	(0.505)	
Bad–Uniform–Good	32.429	0.333	42
	(9.423)	(0.477)	
Uniform–Good–Bad	33.333	0.205	39
	(10.103)	(0.409)	
Total	33.411	0.378	275
	(9.685)	(0.486)	

Notes: The table shows averages per sequence. Standard deviations in parentheses.

3 Hypothesis

Let p^* be the true frequency of the high payoff, whose distribution varies between treatments. Since we expect that attitudes towards ambiguity or to complex risk might make participants state different MAPs in the three treatments, we follow Li et al. (2020) and make the following assumptions:

- the utility of outcomes is fixed. We consider $U(£4) = 1$, $U(£1) = 0$, and $U(£2) = 1/3$;[4]

the Uniform treatments were selected, each with equal probability. This increased the payoffs of all participants who had completed the experiment. This error should not have affected decisions, but only which decision was selected for payment. Participants were informed about the error after the experiment.

[4]We set the utility of the safe payoff such that $U(£2) = x \times U(£4) + (1 - x) \times U(£1)$, where $x \in [0, 1]$. This leads to $x^* = 1/3$.

- participants use a probability weighting function because they perceive the tasks to involve complex risks. Similar to Li et al. (2020), we use Prelec's (1998) *compound invariance* function:

$$w(p) = (exp(-(-ln(p))^{\alpha}))^{\beta}$$

- we use $\alpha = 0.65$ and $\beta = 1.0467$, which according to Li et al. (2020) are the most common values for risky probability weighting;

- participants use "forward" evaluation: they consider the three possible outcomes, and take into account their probabilities;

- participants have the following rank-dependent utility function (Schmeidler, 1989):

$$RDU = w(P(£4)) \times 1 + (w(P(£4) + P(£2)) - w(P(£4))) \times (1/3)$$

  where $P(£4)$ is the probability of receiving the high payoff, $P(£2)$ the probability of receiving the safe payoff, and $P(£1)$ the probability of receiving the low payoff.

In this case, the MAPs which maximize participants' utility in the three treatments are: $MAP_G = 7$ ($RDU = 0.628$), $MAP_U = 8$ ($RDU = 0.495$), and $MAP_B = 9$ ($RDU = 0.439$). This leads us to expect the following ordering of MAPs:

Hypothesis 1 *The MAP in the Good treatment (more mass on high values of p^*) is lower than the MAP in the Uniform treatment (a uniform distribution over p^*), which is lower than the MAP in the Bad treatment (more mass on low values*

of p^*).

$$MAP_G < MAP_U < MAP_B \tag{1}$$

We also consider the alternative hypothesis ($MAP_B < MAP_U < MAP_G$), which could be true if instead participants anchor on visual cues of the distributions, such as the mean.

4 Results

First, we present summary statistics for all decisions, by treatment and by decision order. Next, we run non-parametric tests and ordinary least squares regressions to test the hypothesis. P-values for non-parametric tests are from two-sided tests.[5]

Table 3 presents the average MAP by treatment over all decisions and by decision order. This table already suggests that the hypothesis is not supported by the data, as the average MAP is highest in the Good treatment, followed by the Uniform treatment, followed by the Bad treatment (except for the second decision).

A non-parametric Page's L test confirms this: there is strong evidence that the ordering is the opposite to the one hypothesized ($MAP_B < MAP_U < MAP_G$, p-value < 0.001).[6]

In Table 4 we present results of ordinary least square regressions of MAPs. Model (1) contains as regressors only dummy variables indicating the treatment.

[5]We control for order effects in the regressions.

[6]Page's L test has the null hypothesis that all possible orderings are equally likely. The alternative hypothesis is that a specified order is the increasing order of alternatives. The Stata command is *pagetrend*.

Table 3: Descriptive statistics: MAPs by treatment

	All decisions	First decision	Second decision	Third decision
The Good	9.531	9.571	9.458	9.553
	(2.503)	(2.270)	(2.500)	(2.750)
The Uniform	8.844	8.890	8.368	9.227
	(2.382)	(2.392)	(2.119)	(2.539)
The Bad	8.615	8.093	9.124	8.512
	(2.522)	(2.597)	(2.491)	(2.387)
N	825	275	275	275

Notes: The table shows averages per treatment. Each participant made three decisions in randomized order. Standard deviations in parentheses. Possible answers were integers between 0 and 15.

Model (2) adds age and gender as explanatory variables. Model (3) additionally includes risk attitudes. Model (4) also includes dummy variables for the order in which participants were exposed to treatments. In all models, standard errors are clustered at the individual level.

In all four specifications, participants ask for 0.687 more dark blue sectors on average in the Good treatment compared to the Uniform treatment to be willing to spin the selected wheel (p-value < 0.001 in all specifications). They also ask for 0.229 fewer dark blue sectors in the Bad treatment compared to the Uniform treatment (p-value $= 0.001$ in (4)). More risk loving individuals have lower MAPs (p-value $= 0.04$ in (4)). The only sequence order which differs significantly from the baseline (Good–Uniform–Bad) is Good–Bad–Uniform: MAPs are significantly higher than in the baseline. The first decision (which is Good in the baseline and in Good–Bad–Uniform) differs significantly between these sequences. Since at that point the sequence of events and information participants faced in the two treatments was identical, this difference cannot be a treatment effect or an order

effect.[7]

Result 1. *Participants set the lowest requirement to be willing to take a randomly drawn lottery in the Bad treatment, followed by the Uniform treatment, followed by the Good treatment.*

We speculated that such an ordering of MAPs is possible if individuals anchor on visual cues offered by the distributions. If this were true, then the effects should be reduced if we control for the individual anchoring score (Cheek and Norem, 2017) as measured in the post-experimental questionnaire. This is however not the case. Should I put "not reported" or include it somewhere in the appendix?

5 Discussion

In this note, we wanted to test a necessary assumption for the way betrayal aversion has been elicited in the past to be incentive compatible. This assumption is that the underlying distribution in two treatments which are contrasted to isolate betrayal aversion does not influence behavior. If these underlying distributions do have an influence, then, under certain assumptions, the premium identified as betrayal aversion could actually be due to these differences (Li et al., 2020).

We remove the social/strategic aspects of the game and exogenously manipulate underlying distributions in three treatments. Two of these treatments aim to emulate plausible distributions imagined by subjects in studies on betrayal aversion.[8]

[7] In a robustness check, we reran the regressions separately for each ordering. The signs of the effects are the same for each ordering as for the pooled sample, even if some effects do not reach significance in these smaller samples.

[8] Here I am wondering whether to point out that
- we did take some freedom to imagine the Good distribution–we only calibrated its expected value over the whole distribution. - future research could check whether the way we specified the Good distribution is similar to what trustors believe.

Table 4: Linear regressions on Minimum Acceptable Frequencies

Dependent variable:	Minimum acceptable frequency			
	(1)	(2)	(3)	(4)
The Good	0.687***	0.687***	0.687***	0.687***
	(0.099)	(0.099)	(0.099)	(0.099)
The Bad	−0.229***	−0.229***	−0.229***	−0.229***
	(0.070)	(0.070)	(0.070)	(0.070)
Age		0.005	0.006	0.006
		(0.013)	(0.013)	(0.014)
Male		−0.047	0.030	−0.029
		(0.286)	(0.284)	(0.283)
Risk attitudes (0–10)			−0.172**	−0.152**
			(0.074)	(0.074)
Sequence				
Uniform–Bad–Good				0.490
				(0.408)
Bad–Good–Uniform				−0.008
				(0.434)
Good–Bad–Uniform				1.173***
				(0.412)
Bad–Uniform–Good				−0.150
				(0.434)
Uniform–Good–Bad				0.469
				(0.433)
Constant	8.844***	8.696***	9.520***	9.066***
	(0.144)	(0.460)	(0.593)	(0.627)
N	825	825	825	825

Notes: Standard errors clustered at the individual level in parentheses. The baseline treatment is the Uniform distribution. The baseline sequence is Good–Uniform–Bad. Risk attitudes are measured on a 0–10 scale, where 0 is very risk averse and 10 is very risk loving.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We find a difference in behavior between treatments, but of the opposite sign to our expectation.

This result however is consistent with theories of reference-dependent preferences which predict that individuals will be more risk loving when endowed with more risky options. Since our experiment was not meant to disentangle between competing theories, several of them could explain our results—for instance, the one in Kőszegi and Rabin (2006); Kőszegi and Rabin (2007) or the one of Wenner (2015). These theories state that expectations (which we manipulated exogenously) act as reference points. Modifying expectations modifies the gain-loss component of the utility function, such that higher expectations may make the same outcome less desirable. Alternatively, changing expectations could directly affect consumption utility: if one derives self-image utility from one's consumption, a change in expectations could change which goods are more desirable and thus, offer a boost in self-image (Strahilevitz and Loewenstein, 1998; Marzilli Ericson and Fuster, 2011). In a treatment with better options overall, it is plausible that fewer options increase one's status.

Advantages/contributions of our paper:

– we contribute to the experimental research on the influence of reference dependence on valuation with induced probability beliefs over a receiving a certain amount versus receiving a lottery with fixed payoffs, but varying probabilities for the payoffs;

– we do not endow subjects with either good—though there is some evidence that the fact that we use a visual presentation of the lottery makes it more salient, and thus more likely to be chosen as a reference point;

– since the goods we use are monetary (a fixed amount and a lottery with fixed

payoffs) and since we vary the distribution of probabilities of receiving either good instead of the distribution of prices, we believe subjects have no reason to infer different market values of the goods across treatments;

References

Bohnet, I. and Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, 55(4):467–484.

Cheek, N. N. and Norem, J. K. (2017). Holistic thinkers anchor less: Exploring the roles of self-construal and thinking styles in anchoring susceptibility. *Personality and Individual Differences*, 115:174–176.

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.

Engelmann, D., Friedrichsen, J., van Veldhuizen, R., Vorjohann, P., and Winter, J. (2021). Decomposing trust. Personal correspondence.

Fetchenhauer, D. and Dunning, D. (2012). Betrayal aversion versus principled trustfulness—how to explain risk avoidance and risky choices in trust games. *Journal of Economic Behavior & Organization*, 81(2):534–541.

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

Horowitz, J. K. (2006). The Becker–DeGroot–Marschak mechanism is not necessarily incentive compatible, even for non-random goods. *Economics Letters*, 93(1):6–11.

Karni, E. and Safra, Z. (1987). "Preference reversal" and the observability of preferences by experimental methods. *Econometrica*, 55(3):675–685.

Kőszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. 121(4):1133–1165.

Kőszegi, B. and Rabin, M. (2007). Reference-dependent risk attitudes. *American Economic Review*, 97(4):1047–1073.

Li, C., Turmunkh, U., and Wakker, P. P. (2020). Social and strategic ambiguity versus betrayal aversion. *Games and Economic Behavior*, 123:272–287.

Marzilli Ericson, K. M. and Fuster, A. (2011). Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments. *The Quarterly Journal of Economics*, 126(4):1879–1907.

Polipciuc, M. (2021). Group identity and betrayal: decomposing trust. Working paper.

Polipciuc, M. and Strobel, M. (2020). Betrayal aversion with and without a motive. Working paper.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 66(3):497–527.

Quercia, S. (2016). Eliciting and measuring betrayal aversion using the BDM mechanism. *Journal of the Economic Science Association*, 2(1):48–59.

Scheier, M. F., Carver, C. S., and Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology*, 67(6):1063–1078.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica*, 57(3):571.

Stephenson, M. T., Hoyle, R. H., Palmgreen, P., and Slater, M. D. (2003). Brief measures of sensation seeking for screening and large-scale surveys. *Drug and Alcohol Dependence*, 72(3):279–286.

Strahilevitz, M. A. and Loewenstein, G. (1998). The effect of ownership history on the valuation of objects. *Journal of Consumer Research*, 25(3):276–289.

Thomson, K. S. and Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, 11(1):99–113.

Tymula, A., Woelbert, E., and Glimcher, P. (2016). Flexible valuations for consumer goods as measured by the Becker–DeGroot–Marschak mechanism. *Journal of Neuroscience, Psychology, and Economics*, 9(2):65–77.

Wenner, L. M. (2015). Expected prices as reference points—theory and experiments. 75:60–79.