NCER Working Paper Series

Within-subject Intra– and Inter-method consistency of two experimental risk attitude elicitation methods

Uwe Dulleck
Jacob Fell
Jonas Fooken

Working Paper #74
October 2011

Within-subject Intra- and Inter-method consistency of two experimental risk attitude elicitation methods

Uwe Dulleck[1], Jacob Fell[2] and Jonas Fooken[3]

September 27, 2011

Abstract

We compare the consistency of choices in two methods to used elicit risk preferences on an aggregate as well as on an individual level. We asked subjects to choose twice from a list of nine decision between two lotteries, as introduced by Holt and Laury (2002, 2005) alternating with nine decisions using the budget approach introduced by Andreoni and Harbaugh (2009). We find that while on an aggregate (subject pool) level the results are (roughly) consistent, on an individual (within-subject) level, behavior is far from consistent. Within each method as well as across methods we observe low correlations. This again questions the reliability of experimental risk elicitation measures and the ability to use results from such methods to control for the risk aversion of subjects when explaining effects in other experimental games.

Keywords risk preferences · laboratory experiment · elicitation methods · subject heterogeneity
JEL classification C91 · D81

[1]Queensland University of Technology, School of Economics and Finance, 2, George Street, 4001 Brisbane, QLD, Australia, Email: uwe.dulleck@qut.edu.au

[2]The Commonwealth Grants Commission,86-88 Northbourne Ave, 2612 Braddon, ACT, Australia, Email: jacob.fell@gmail.com

[3]Queensland University of Technology, School of Economics and Finance, 2, George Street, 4001 Brisbane, QLD, Australia, Email: j.fooken@qut.edu.au (Corresponding author)

1 Introduction

Measuring and controlling for risk aversion in the laboratory is a commonplace in many economic experiments. However, while risk aversion is a relatively straight forward theoretical concept - where its extent is easy to determine for a given utility function - testing for risk aversion in economic experiments is less trivial. A large body of the literature suggests methods for as well as discusses problems with the elicitation of risk attitudes: The interpretation of empirical results, in even the most carefully designed experiments, is not nearly as clean and straight forward as theory would predict. Accordingly, to explain behavior, contributions in the literature rely on notions of stochastic elements in individual choices (e.g. Loomes and Sugden, 1995, 1998; Loomes et al., 2002), model the effect of interdependence between choice options presented (Starmer and Sugden, 1993) or capture the idea of heterogeneity between (possibly types of) players (Ballinger and Wilcox, 1997). However, this literature - despite its insightful considerations - does not provide an easily applicable toolkit for the elicitation of risk attitudes in a laboratory environment to overcome the problem of inconsistencies observed in choice patterns of experimental subjects.

Harrison and Rutström (2008) address this issue of handling experimental data in a survey that reviews different risk elicitation methods and discusses ways to estimate risk attitudes. While this review compares different elicitation methods and discusses specific characteristics of the methods, it only compares (cross-sectional) aggregate information and does not compare differences in elicitation methods on an individual level. One reason for this might be that several studies have found that individual as well as aggregate differences in risk attitude measurements depend on different elicitation methods. Isaac and James (2000) compare implied risk attitudes of 34 subjects that resulted out of choices made in a first price auction and by using the Becker-DeGroot-Marschak (BDM) procedure,[1] finding that experimental choices imply different risk attitudes. I.e., experimental decisions for the same individuals using the two methods cannot be captured by the same or similar utility functions. Their results indicate that the two methods do not just serve as pure shifters of risk aversion within individuals: ranked correlations (across individuals) are only around 39%. From a practical viewpoint, this questions the usefulness of the methods to control for previously determined risk aversion in (other) experimental games. A number of studies investigated measures of risk aversion within individuals, generally

[1] This procedure is used to determine certainty equivalents to a given lottery.

finding that risk attitudes were not stable within individuals in experimental settings: Berg et al. (2005) found that implied risk attitudes depend on whether individual decisions are measured using auctions for a risky or a riskless asset. Hey et al. (2009) compare willingness-to-pay, willingness-to-accept, BDM measures and choices over pairwise lotteries. They find inconsistencies and in some cases even negative correlations between results of the different methods within individuals. Anderson and Mellor (2009) compare results of the method developed by Holt and Laury (HL, 2002) and survey results on gambles (over job and investment choices), finding that except for a small fraction of *superconsistent* ("consistently consistent", p.152) decision makers, the methods did not provide consistent within-individual estimates of risk attitudes. Comparing HL results and decisions over a decision they refer to as the "Deal or No Deal game"(named after a popular TV show), Deck et al. (2008) find that decisions are not consistent and conclude that one elicitation method is treated as an investment (HL), while the other as a gambling decision. Harrison et al. (2005b) found that risk attitudes measured using HL were unstable over a period of six months. Lönnqvist et al. (2011) also look at intertemporal stability using HL and a survey; their results indicate that the assumption of stability is problematic and that the predictive power of implied risk attitudes based on HL and decisions in the trust game is low.

Each of these studies compare the results from one risk elicitation method with the results from another choice setting (like an auction, a trust game or a survey) from which risk attitudes can be inferred. Our approach differs from this literature by comparing the results of two risk elicitation methods (each applied twice) to measure within-subject stability over a short time frame as well as cross-method consistency. Closest to our study is Dave et al. (2010, in a study on a cross-section of the Canadian population), who also compare the results of two methods, i.e. HL and an approach by Eckel and Grossman (EG, 2002). They find that implied risk attitudes of the two methods differ (in the EG method more individuals are risk neutral) and that HL leads to more inconstistent choices, particularly among individuals with lower mathematical skills. This literature indicates that individual risk attitudes are not stable, and should be treated with caution when interpreting experimental results. While our understanding of risk aversion rests upon the notion of an underlying utility function as an individual-based concept, experimental evidence does not easily align with a utility function that is independent of the method used and consistent over time. Following the above research we take the idea of cross-method comparison as our starting point and try to compare two risk elicitation methods. However, in order to

have a measure of comparison at hand, we compare the consistency across methods with their consistency within each method. Furthermore, we use two risk elicitation methods for which the decision variable is the same, i.e. an optimal probability over gains, reducing the potential for a bias caused by a different decision variable.

In our analysis we find (a) that both methods give a divergent picture of the overall risk attitude of the subject pool, (b) that consistency of individual decisions throughout the experiment is limited for both methods and (c) that individual-level consistency decreases further when comparing the two methods. These results confirm outcomes of prior research and call into question in how far experimental results on risk attitudes can be used for more than very general statements about decisions of the whole subject pool. The observation is further aggravated considering that the internal consistency is not much better within than across methods, i.e. the problem does not only seem to be that measures depend on framing.

1.1 Desirable characteristics of a risk elicitation method

What are desirable characteristics of a risk elicitation method? Knowing that not all individuals are identical, methods need to allow for heterogeneity of risk attitudes in the population of experimental participants. Preferably the choices observed using a method provide information about individual subjects partaking in an experiment rather than a general statement about a group of participants in a (laboratory) study. For example, it would be desirable if a method would allow researchers to classify participants into groups of risk averse, risk neutral and risk loving individuals. A "perfect"elicitation method would even enable researchers to make predictions about acceptable risk premia of individuals for given choice options and (or) to estimate reliable coefficients of utility functions for each individual. We will consider these desired but pragmatic criteria when evaluating our results.

1.2 The Methods

While these desirable characteristics of risk elicitation methods are relatively straightforward, they are usually not directly addressed in experimental studies. Many studies are based on a theoretical framework of risk aversion assuming stable individual risk attitudes. In contrast to this fixation on the individual in the theoretical world, evaluations of experimental results often only look at aggregate measures over a subject pool. One reason for this might be that robust individual-level measures of risk attitudes

are difficult to find with the given methods. To see whether this is the case we look at two methods that start from a theoretically similar idea of utility functions, i.e. utility with constant relative risk aversion (CRRA), in which the choice variable is similar for experimental participants, they choose probabilities. Both have been designed to elicit risk attitudes in the laboratory, i.e. they are somewhat laboratory-artificial and do not directly relate to real-life choice problems. Furthermore, we incentivize both methods such that they would yield the same expected value for a risk-neutral decision maker. We replicate some results of the original studies and compare their coefficient estimates on an aggregate, full-sample level as well as within individuals. We alternate the order of the methods and each subject makes the choices in both methods twice, alternating between the two methods for each subject. This allows us to compare the results within each method as well as comparing within-individual choices across methods.

The first of these two methods is the one used by HL, which uses a menu of lotteries (or multiple price list, MPL) with changing probabilities over constant pairs of outcomes. For both options one outcome is higher than the other and in all cases in the first option the difference between outcomes is small while for the second option the difference is large. In total ten decisions are presented in the order where the probability of receiving the higher payoff is increasing both for the first and the second option. As the variance of outcomes for the first option is always lower, the number of choices for the first option gives a measure of risk aversion. Furthermore, any utility concept that is monotonic in probabilities given constant outcomes yields one switching point from the first to the second option. In their study, HL find that subjects are generally risk averse and that risk aversion increases with the size of the stakes, a statement they refined in a second study (Holt and Laury, 2005) after a comment by Harrison et al. (2005a). Since it's publication, HL's method has been used in several studies as it allows to determine risk premia that experimental participants are willing to pay for experimental lotteries. HL can also be used to infer CRRA coefficients. Despite the (theoretically) straightforward design of the method and its popularity in the literature, HL is not flawless in the sense that it often leads to inconsistencies, i.e. multiple switching points, which can be interpreted as relatively broad bandwidths of possible risk premia - including ambiguity on whether individuals are risk loving, risk neutral or risk averse. Furthermore, while it is possible to test whether decisions revealed using the HL approach are compatible with a certain utility function with known parameters, HL's method cannot (easily or non-numerically) be used to determine or calibrate such a utility function of individuals. This, in comparison, is possible using the

second method we employ. In a more recently proposed method Andreoni and Harbaugh (AH, 2009) let individuals allocate a budget in each decision over the probability of winning and the gain in case of winning, whereas the alternative to winning is always an outcome of 0. The AH set-up allows a direct calculation of a CRRA coefficient from every decision taken, which increases the number of observations that can be collected on one individual during the experiment.[2] We use AH as the method of comparison in our experiment.

2 Experimental Design and Procedures

We use a within-subject design of individuals that make choices based on the risk elicitation methods introduced by HL and AH. We analyze decisions of 78 experimental participants from a regular student population throughout 7 sessions. Participants were recruited online from the experimental subject pool at the Queensland University of Technology using ORSEE (Greiner, 2004) and through announcements in tutorials. Some participants were also recruited in common places at the university in personal communication; however, when asking students in person for participating in the experiment, the same information was used for recruitement, including the organizer (researchers at the School of Economics and Finance), average earnings (around 20 Australian dollars) and time estimated to complete the experiment (around 30 minutes). It was also pointed out to the students that there would be no minimum payment for participating in the experiment and participants were also motivated by potential gains of up to 150 Australian dollars. It is worth noting that this recruitement of asking students personally to participate was somewhat less controlled than common in many economic experiments. However, as we were interested in within-subject comparisons and were still drawing from a relatively homogeneous student population, this was of minor concern for this study. The risk elicitation methods were implemented in a computer laboratory using a custom-made, java-based software. Upon arrival in the laboratory participants were seated at computers, were asked to work through experimental instructions and start the experiment. Instructions included examples of how to make choices in the experiment and two test questions for each risk elicitation method. Further help by the experimenter was available upon re-

[2] In comparison, the MPL of the HL method requires an individual to make several decisions to infer bounds for such a coefficient. Consequently, more choices are necessary to have one risk attitude observation for an individual.

quest of participants. When participants had passed the test questions, they started the experiment, going through two rounds of 9 choices for each risk elicitation method. The order of the risk elicitation methods was switched for about half of our experimental sessions (we did not find significant order effects across participant's decisions depending on the order of the methods). After completing the experiment, participant's were given the opportunity to change their earlier decisions for the experimental round that would be chosen for final payoff. Finally, participants were given a questionnaire that asked for some demographic information and student status. After students had finished the questionnaire they were paid and could leave the computer laboratory.

2.1 Holt and Laury's Method

In the design of the risk elicitation tasks we followed the design chosen by HL and AH closely. Therefore, we just outline our approach briefly here and refer to the original papers for full detail. Furthermore, screenshots can be found in the appendix. For the HL method, participants were able to see a MPL and were asked to make choices separately for each row between a pair of lotteries. For each further decision row down, the probability mass on the higher payoff increased by 10%, making the safer option A (i.e. the option with a lower variance in payoffs) less attractive. Contrary to HL we, however, left out the 100% option of the higher payoff (i.e., we did not include a choice in which one option is dominated), which reduced our number of choices from 10 to 9 in each round. Over the two rounds we played one set-up in which participants played over high (a dollar gamble between 10 and 8 vs. a gamble between 19.25 and 0.5) and one in which they played over slightly lower stakes (8 and 6.4 vs. 15.4 and 0.4). Both payoffs are closer to the low than to the high payoff treatment in HL. The high stake set-up scaled up payoffs and therefore implied higher risk premia for choosing the more secure option. However, the estimated bounds for CRRA coefficients remain at the same number of safe choices. Under the assumption that a participant has a constant CRRA coefficient the same optimal switching (point) probability in both rounds is chosen. Table 1 provides an example of the low-payoff set-up.

2.2 Andreoni and Harbaugh's Method

Similarly, we implemented the AH risk elicitation method following the original approach. However, we did not use gambles over negative amounts, but

Table 1: Multiple price list design by Holt and Laury

Option A					Option B			
p	*X*	*1-p*	*Y*		*p*	*X*	*1-p*	*Y*
0.1	8	0.9	6.4		0.1	15.4	0.9	0.4
0.2	8	0.8	6.4		0.2	15.4	0.8	0.4
0.3	8	0.7	6.4		0.3	15.4	0.7	0.4
0.4	8	0.6	6.4		0.4	15.4	0.6	0.4
0.5	8	0.5	6.4		0.5	15.4	0.5	0.4
0.6	8	0.4	6.4		0.6	15.4	0.4	0.4
0.7	8	0.3	6.4		0.7	15.4	0.3	0.4
0.8	8	0.2	6.4		0.8	15.4	0.2	0.4
0.9	8	0.1	6.4		0.9	15.4	0.1	0.4

Individuals are asked to chose between Option A or B for each row.

restricted our experiment to positive gambles only. In the AH method participants were able to see the probability of winning a certain amount or receiving zero otherwise. The probability of winning was illustrated as a green shaded area in a pie chart. The amount received in case of winning was illustrated as a green shaded area in a bar chart that filled up representing the higher gain in case of winning. Participants were able to change the probability of winning by moving a slider whereas every extra percent of winning meant that the potential gain in the gamble would be reduced by a (constant) price. I.e., increasing the probability of winning was costly in terms of the amount won. While the gamble was graphically represented on the computer screen, the probability of winning and the amount to be won were also stated as numbers on the screen. Table 2 shows the maximum amount that could be won with probability zero in a round and the price of one extra percentage probability of winning. These combinations were each presented to participants twice.

Finally, the design of the payoff structure for the two methods was designed such that the expected gain from the 18 decisions in each method for a risk neutral decision maker was the same across the two risk elicitation methods in order to keep both set-ups as comparable as possible.

Table 2: Pairs of maximum gain and cost of probability

Round	A	B	C	D	E	F	G	H	I
Maximum gain (μ)	27.3	56	172	88	49.4	39.2	54.5	207	116
Price of 1 extra percent of winning probability (price(p))	0.28	1.17	10.75	2.75	0.77	0.41	0.68	8.62	2.42

Individuals chose over p, facing the constraint that they will receive $\mu - p \cdot price(p)$ with probability p.

3 Analysis of replication and of aggregate decisions

In a first step we replicated some of the (central) results in the approaches by HL and AH that were relevant for our comparison. Both papers considered deriving parameter estimates for a CRRA utility function of the form $U(x) = \frac{x^{1-r}}{1-r}$, as introduced in HL or similarly $U(x) = x^{\alpha}$ as in AH. In both methods the probability chosen was the main choice variable of interest for the analysis. For this utility function, HL grouped experimental decision makers into categories of individuals with a certain risk attitude, based on their coefficient r. Although the method used by HL does not allow to directly calculate such a coefficient, bounds of it can be determined by looking at the switching points from more risky to less risky choices. These bounds are, however, difficult to identify if individuals have more than one switching point. Dealing with these issues, HL counted the number of safe choices that an individual had made and categorized individuals into categories that this number of safe choices would have implied if they had only a single switching point (SSP). Table 3 reports our replicated results for two payoff set-ups comparable to the low set-up of HL, as well as the original results in HL in their two treatments with low and high monetary payoffs. The last column contains the empirical distribution of CRRA coefficients based on our AH data to allow a comparison.

The AH risk elicitation method allows for a straight forward calculation of CRRA coefficients under the functional form as described above; we do this for each decision that experimental participants take and report the distribution of all the decisions by all participants based on the implied

Table 3: Overall distribution of risk attitudes

Risk attitude	Number of safe choices		HL (repl.) (1)	HL (repl.) (2)	HL (2002) (3)	HL (2002) (4)	AH (5)
Highly risk loving	0-1	$r < -.95$	1%	1%	1%	1%	5%
Very risk loving	2	$-.95 < r < -.49$	0%	7%	1%	1%	2%
Risk loving	3	$-.49 < r < -.15$	8%	5%	6%	4%	6%
Risk neutral	4	$-.15 < r < .15$	29%	21%	26%	13%	61%
Slightly risk averse	5	$.15 < r < .41$	17%	23%	26%	19%	11%
Risk averse	6	$.41 < r < .68$	22%	19%	23%	23%	9%
Very risk averse	7	$.68 < r < .97$	10%	22%	13%	22%	4%
Highly risk averse	8	$.97 < r < 1.37$	4%	4%	3%	11%	2%
Stay in bed	9-10	$1.3 < r$	9%	4%	1%	6%	0%

The table shows the share of decisions that would be classified in risk categories as proposed by HL. We include our replicated HL results (1) and (2), the results from HL's original (2002) paper (3) and (4), as well as results implied by our data about the AH method (5). Our stakes are higher in (2) than in (1), but both correspond to the low stakes treatment (3) in HL's original approach, as they are significantly lower than in HL's high stakes treatment (4).

r-coefficient.[3] We do not replicate the full analysis by AH, who answer five questions on expected utility. Instead we focus on whether using a CRRA framework with a simple utility function as characterized before is reasonable. We confirm their regression results over all decisions showing that budget allocations of the winning probability and the winning price are approximately constant over the size of winning stakes. This indicates that CRRA is a reasonable assumption. We do not continue replicating further results reported in AH, as our main aim is to compare the two methods by HL and AH. Doing so, we find that the classification in terms of risk attitudes of our subject pool when using the HL method follows a similar distribution to the one reported by HL in their original contribution. Furthermore, we can generally identify a noticeable degree of risk aversion in our subject pool and also find a tendency of (slightly) increasing risk aversion when the stakes over which the lotteries are played increase. We also find that coefficients calculated using AH's method provide results that indicate a higher number

[3] Given that in AH $U(x;p) = p \cdot x^{\alpha}$ and $x = \mu - \text{price}(p) \cdot p$, $\text{argmax}_p\, U(x;p) = \text{argmax}_p\, ln[U(x;p)] = \frac{\mu}{\text{price}(p)\cdot(\alpha+1)}$, where μ is the maximum gain in a period that can be chosen with a corresponding probability of zero. As μ and price(p) are known, for a chosen p we can calculate $\alpha = \frac{\mu}{\text{price}(p)\cdot p} - 1$, which can be transformed to $r = \alpha - 1$.

of risk neutral choices compared to results in HL, some risk averse choices as well as some decisions that are risk loving. Analyzing the total distribution of the results indicates that the two methods, despite drawing on a very similar notion of utility functions and both being theoretically legitimate risk elicitation procedures, do not provide with the same result. I.e., the average risk attitude in HL is between slightly risk averse and risk averse, while the average decision in AH is risk neutral (with a tendency towards risk aversion). This is true despite the fact that the expected monetary payoff from participating in the experiment is the same across methods.

4 Analysis of individual decisions

In order to get a better understanding of these differences in the results, we try to analyze the decisions of our participants on a within-subject basis. That is, since all of our participants made 18 decisions in each method, we can analyze in how far each individual decided consistently within and across the two methods.

4.1 Internal consistency of the methods

In a first step we analyzed in how far individual participants made consistent decisions within one risk elicitation method. We did so to have a benchmark of comparability when looking across methods. For this, we used correlations of individual decisions over the two rounds. For the HL method, the number of safe choices made in the first and the second period, which were used to calculate CRRA coefficients as shown in Table 3 gave a correlation of 55% and a ranked (Spearman's ρ) correlation of 62%. We also considered a second way to measure the degree of risk aversion for which we did not assume that participants have a clearly determinable SSP, but calculated the average risk premium within their farthest switching points. (This corresponds to an approach described by Andersen et al. 2006.) These averages were correlated at a level of 68% over the two rounds of HL. Figure 4 in the appendix also provides a picture of the dispersion of the difference between safe choices in the first round (over lower stakes) and the second round (over higher stakes), indicating that there is a slight shift towards risk aversion, but that it is not a one-directional shift.

As in the HL method the idea of a SSP from less to more risky options is important, we also looked at whether assuming the general prevalence of SSP was reasonable for our sample, and how many of the players with SSP consistently chose the same number of safe choices over the two rounds.

From our 78 participants in the experiment, 48 players had a SSP in both rounds of the HL method.[4] Of these, 22 chose the same number of safe choices in both periods, although one of them changed the decision when being able to reconsider their choice at the end of the experiment. Of the 22 (HL-consistent) individuals, 10 participants were in the risk neutral category as introduced above and 12 were either risk averse or risk loving. Finally, we also looked at whether participants wanted to change their decisions in the round that was played for final payout. However, in the HL set-up, only 8 out of 78 participants wanted to change their decision indicating that most participants had already made their best-informed choice before; one of these 8 increased the number of safe choices, while all others increased the number of risky choices.

To analyze the internal consistency in the AH set-up, we similarly first looked at correlations between decisions of individuals made between the rounds. For this purpose we calculated implied CRRA α-coefficients for each decision as described in footnote 3. These coefficients showed correlations that ranged between 15% and 60% for the same lottery (i.e. the same choice over a corresponding maximum gain and price of an extra probability of winning) over the two rounds. Ranked correlations were between 30% and 57% across individuals.[5] There was, however, no apparent relationship between the stake of the lottery and the correlation between the two rounds; that is, it was not clear how to identify which factors led to higher consistency over the rounds. Figures 1 a and b illustrate these correlations for each round.

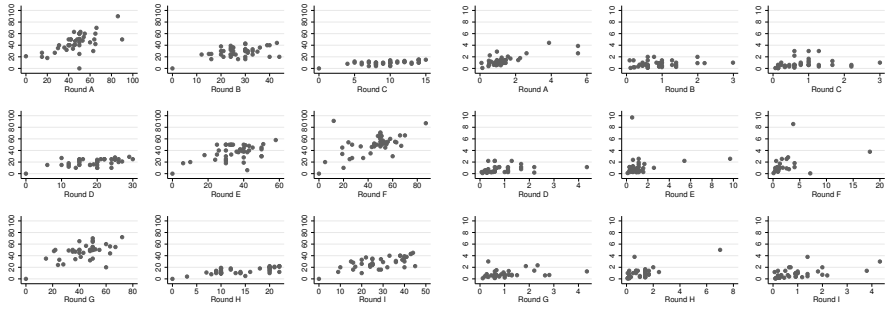
In a second step we therefore tried to find an individual aggregate for the CRRA coefficient over the different rounds. We did so by averaging the coefficients for each individual over each round. In order to find out if such an aggregation was appropriate, we tested for whether there was a positive or negative relationship between the maximum gain and the implied CRRA coefficient. While we found that it did so for some participants, it did not for most individuals. To get a better understanding of this ambiguous result, we

[4]The rate of individuals that had more than one switching point in our study is comparatively high, at least when compared to the other studies mentioned in the introduction that used the HL method; they reported non-consistent individuals and non-SSP individuals with shares between 2% and 9% of the sample. Our single round rate is higher than this and our rate is further increased as we played the HL game over two rounds.

[5]The correlations of α-coefficients understates the correlation of probabilities chosen over the 2 rounds, as small differences in probabilities chosen over the rounds that are far away from the risk neutral choice optimum are amplified. Hence, for comparison we also looked at the probabilities chosen over the rounds; these are correlated between 40% and 63%.

Figure 1: Correlations of decisions over two rounds of AH

(a) Correlations of probabilities (b) Correlations of α-coefficients

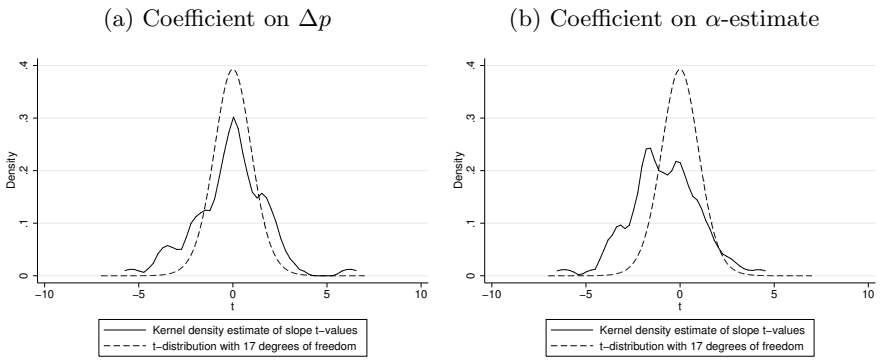


compared the t-values for the estimated β -coefficients in regressions of the maximum gain on deviations from risk neutral probabilities and regressions of the maximum gain on the individual CRRA coefficient with what would be expected under a t-density for our estimation.[6] We find that the difference between the two distributions is statistically insignificant (t=-0.97) for the deviation from the risk neutral optimum and marginally significant for the α-coefficients (t=-1.95). We take this as a fair approximation to treat the CRRA coefficient for AH as constant over the size of stakes and proceed by averaging them. Figure 2 a and b illustrate the two comparisons between the distribution of the estimated individual β-coefficients and a t-distribution.

Having done this aggregation, we compared (round) average α-values over the two rounds; they showed a correlation of 70% by individual and a ranked correlation of 72%. In order to get a better picture of robustness of the CRRA coefficients, we also looked at whether participants changed their decisions when being informed that a certain round would be selected for final payoff. The result showed that – comparatively to the HL method – many participants (a total of 27) changed their choices. Furthermore, the percentage change of those individuals that revised their decisions was noticeable; on average, participants that changed their choices moved 12% towards safer choices and absolute changes were 30%. Finally, we inves-

[6]Practically, for each participants we ran an ordinary least squares regression of the form $y_i = \phi_i + \beta_i \cdot \mu_i + \epsilon_i$, using the 18 observations of each participant i, saved all β_i estimates and compared the distribution of the estimates to a t-distribution (with 17 degrees of freedom). For y_i we first used $\Delta p = p_{opt} - p_i$, where p_{opt} was the optimal probability chosen by a risk neutral individual and p_i the probability chosen by individual i. Secondly, we also did the same analysis with $y_i = \alpha_i$.

Figure 2: Comparison between the distribution of slope parameters



tigated in how far using average CRRA coefficients derived using the AH method allowed us to reliably classify participants into broad categories of risk averse, risk neutral and risk loving individuals. We therefore tested whether the average CRRA coefficient α was significantly different from one (or $r \neq 0$ using HL's terminology) using confidence intervals of 2 within-subject standard deviations. We found that only for 5 participants out of the 78 the CRRA coefficient α was significantly different from one; i.e. from our estimates these 5 participants were risk averse and all other participants were approximately risk neutral.[7]

4.2 Comparison across methods

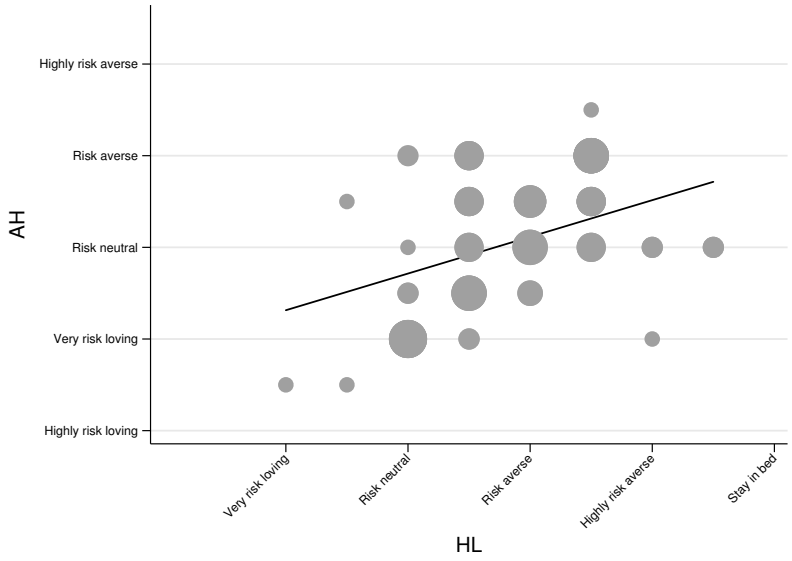
Finally, our data allows us to compare the two risk elicitation methods on a within-individual basis. One way to do so is trying to make predictions based on one method of how an individual would have made decisions in the other method. Following this rationale, we used the average risk aversion coefficient derived using the AH elicitation procedure to predict how an individual with this parameter would have decided in the HL framework.

We found that this would have predicted 76% and 75% of decisions in the two rounds of the HL method, respectively. However, in this comparison any individual that has multiple switching points (MSP) will have some incorrect predictions, despite the fact that the estimate used is not incorrect. To alleviate this effect, we looked at individuals with SSP only, which showed

[7] Main reason for this is that for almost all participants the estimated standard deviation on α is **s.d.**≥ 0.3, as can be seen in Figure 5 in the appendix.

83% and 82% correct predictions over the two periods, respectively. However, while this might seem like a high level of comparability, these numbers have to be read with care, as instead of using the AH aggregate just guessing all individuals to be risk neutral (although this is not what our aggregate analysis suggested) would have predicted choices made by individuals under the HL method equally well. We therefore reverted to the categorization of participants into groups of people with different risk attitudes as in Table 3. We allocated individuals into these risk categories according to the two methods. Using this approach, 10% of participants were grouped into the same risk attitude category by both methods. Reasons for this are that the AH method (on average) shifts risk averse individuals in the risk neutral category (a shift of about 27%); however, the shift is not only in one direction (the average absolute shift is about 33%). Consequently, when looking at the ranked correlation on allocations to risk categories the result is that the two methods are (rank) correlated at about 38%. Figure 3 illustrates this relationship.

Figure 3: Allocation of individuals into risk categories by HL and AH



The number of individuals at a point is indicated by the size of the bubble. The line displays a corresponding linear fit.

5 Conclusion

Using the risk elicitation methods developed by HL and AH, we tested their internal and external consistency across and within individuals. We found that within method correlations of about 60% to 70% of decisions between periods could be established. Comparatively, cross-method predictions and correlations were smaller and could only be established on an aggregate level. Furthermore, the two methods did not necessarily seem to be procedurally invariant, both over the full subject pool (as visible in Table 3), as well as on an individual level. This seems undesirable considering that *a priori* one would have guessed that the two methods would yield similar results and it seems difficult to determine a better method *ex post*. Also, this puzzle does not seem to be rooted in the decision variable, as in both methods individuals chose over probabilities. Part of the reason for this procedural variance is that the AH method shifts decisions toward risk neutrality. Ranked correlations are hence higher, at around 38% between the methods, which is surprisingly close to what Isaac and James (2000) found in their paper comparing methods. As most of the literature before, we read these individual-based cross-method correlations as (somewhat unsatisfactory) low; however, these low correlations are also due to the low consistency of decisions for even the same procedure, which can be observed in our within-method benchmark. Furthermore, we do not find many *super-consistent* individuals as, for example Anderson and Mellor (2009) found in their study; i.e. at least in our subject pool individual inconsistencies are an almost universal problem. While we have no clear means to determine which of the two methods is the correct or superior one, from our results we can evaluate in how far the desirable characteristics mentioned in the beginning of the paper are met by the two methods. First of all, in the aggregate both methods allow for making statements about the overall risk attitudes of the subject pool and we would conclude that the subject pool is on average (moderately) risk averse. This confirms a general idea that on average individuals prefer safe options to gambles. This conclusion is true for both methods, although results using the method by AH would suggest that most individuals are more centered around risk neutrality. However, while both methods are able to make a statement about the risk attitude of the overall subject pool, it seems difficult to be able to reliably infer the risk attitude of an individual from the methods. While the HL method was more consistent over the two rounds than the AH method, for both methods it seems problematic to clearly identify the risk attitude of an individual. I.e., over all 18 decisions in the AH method, it was not possible to identify more than

5 of 78 participants having a CRRA coefficient significantly different from $\alpha = 0$, although the overall picture suggests that there is risk aversion in the population. Upon first sight the HL method performs better on this ground, but still only 22 participants make consistent decisions, and again for most participants it is unclear whether they are risk averse, risk neutral or risk seeking. This conclusion remains despite the fact we only repeated the HL task over two rounds and one would conjecture that increasing the number of repetitions might lead to more inconsistencies. Furthermore, much analysis of the HL method relies on disregarding or simplifying many inconsistent or mistaken choices that are observable in the data, which might not be advisable, as a study by Jacobson and Petrie (2009) has shown. Finally, it seems that both risk elicitation methods, despite providing some usable aggregate results are not as good as would be desirable in determining individual risk attitudes, which remain ambiguous for most of our participants. Unfortunately, this effect is even severed when adding another risk elicitation method, which shows that estimates are not method invariant. In our study this was true both from a global point of view as well as on an individual level. This is somewhat disappointing considering that risk aversion, based on the notion of individual utility, is essentially an individual-based concept. Without individual consistency of decisions, it is also questionable to what extent HL (or AH) can be used for measuring to control for risk aversion, as it is often done when interpreting other experimental games, for example when analyzing trust or contributions to public goods, to name just two.

References

Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2006). Elicitation using multiple price list formats. *Experimental Economics*, 9(4):383–405.

Anderson, L. R. and Mellor, J. M. (2009). Are risk preferences stable? comparing an experimental measure with a validated survey-based measure. *Journal of Risk and Uncertainty*, 39(2):137–160.

Andreoni, J. and Harbaugh, W. (2009). Unexpected utility: Experimental tests of five key questions about preferences over risk. *Working Paper*.

Ballinger, T. P. and Wilcox, N. T. (1997). Decisions, error and heterogeneity. *The Economic Journal*, pages 1090–1105.

Berg, J., Dickhaut, J., and McCabe, K. (2005). Risk preference instability across institutions: A dilemma. *Proceedings of the National Academy of Sciences of the United States of America*, 102(11):4209.

Dave, C., Eckel, C. C., Johnson, C. A., and Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, pages 1–25.

Deck, C., Lee, J., Reyes, J., and Rosen, C. (2008). Measuring risk attitudes controlling for personality traits. *Working Paper; University of Arkansas, Florida International University*.

Eckel, C. C. and Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, 23(4):281–295.

Greiner, B. (2004). The online recruitment system ORSEE 2.0 - a guide for the organization of experiments in economics. Technical report, University of Cologne, Department of Economics.

Harrison, G. W., Johnson, E., McInnes, M. M., and Rutström, E. E. (2005a). Risk aversion and incentive effects: Comment. *American Economic Review*, pages 897–901.

Harrison, G. W., Johnson, E., McInnes, M. M., and Rutström, E. E. (2005b). Temporal stability of estimates of risk aversion. *Applied Financial Economics Letters*, 1(1):31–35.

Harrison, G. W. and Rutström, E. E. (2008). Risk aversion in the laboratory. In *Risk Aversion in Experiments*, volume 12 of *Research in Experimental Economics*. Emerald Group Publishing.

Hey, J. D., Morone, A., and Schmidt, U. (2009). Noise and bias in eliciting preferences. *Journal of Risk and Uncertainty*, 39(3):213–235.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655.

Holt, C. A. and Laury, S. K. (2005). Risk aversion and incentive effects: New data without order effects. *American Economic Review*, 95(3):902–904.

Isaac, R. M. and James, D. (2000). Just who are you calling risk averse? *Journal of Risk and Uncertainty*, 20(2):177–187.

Jacobson, S. and Petrie, R. (2009). Learning from mistakes: What do inconsistent choices over risk tell us? *Journal of Risk and Uncertainty*, 38(2):143–158.

Lönnqvist, J. E., Verkasalo, M., Walkowitz, G., and Wichardt, P. C. (2011). Measuring individual risk attitudes in the lab: task or ask? an empirical comparison. *SOEP Papers on Multidisciplinary Panel Data Research*.

Loomes, G., Moffatt, P. G., and Sugden, R. (2002). A microeconometric test of alternative stochastic theories of risky choice. *Journal of Risk and Uncertainty*, 24(2):103–130.

Loomes, G. and Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, 39(3-4):641–648.

Loomes, G. and Sugden, R. (1998). Testing different stochastic specificationsof risky choice. *Economica*, 65(260):581–598.

Starmer, C. and Sugden, R. (1993). Testing for juxtaposition and event-splitting effects. *Journal of Risk and Uncertainty*, 6(3):235–254.

A Information on subject pool

Table 4: Summary statistics on experimental participants

Variable		Value
Age	Avg	21.45
	Min	17
	Max	40
	Std	4.04
Gender	Male	50
	Female	28
English Speaker	yes	69
	no	9
Experience with experiments	yes	30
	no	48
Marital Status	Married	4
	In partnership	28
	Single	46
Weekly Income	<$100	33
	$100 - $199	23
	$200 - $299	12
	$300 - $399	5
	$400 - $500	2
	>$500	3
Degree	Business	33
	Engineering	9
	Science	6
	IT	7
	Law	3
	Double	20
Living Situation	Living alone (renting)	7
	Living alone (owning)	1
	Living with partner	10
	Living with your parents	45
	Living in a shared house	12
	Other	3
Cultural backround	Australian	53
	American	1
	Asian	15
	European	6
	African	3

B Figures

Figure 4: Histogram of individual differences between the number of safe choices over the two rounds in HL

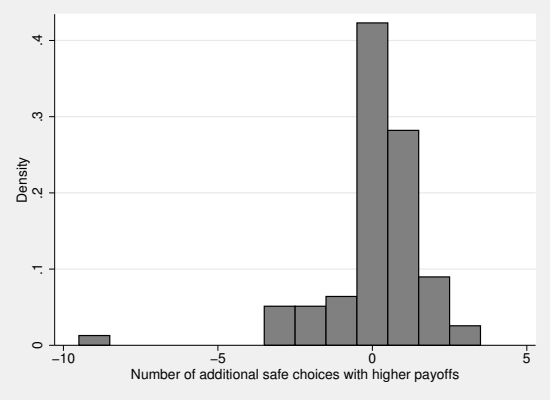
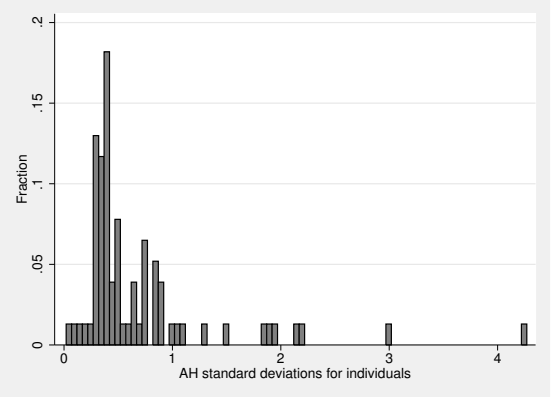


Figure 5: Histogram of α-standard deviations of experimental participants in AH



C Examples of experimental screens

Figure 6: Screenshot from our experiment using HL

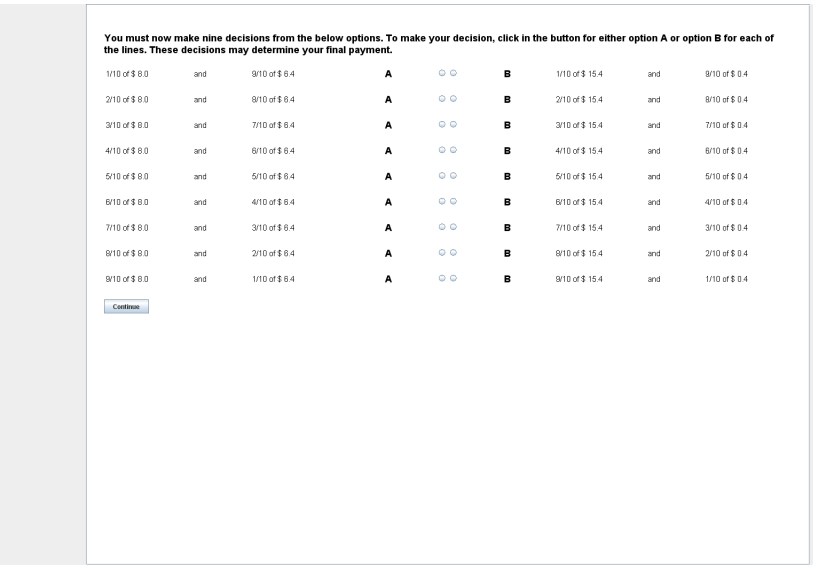


Figure 7: Screenshot from our experiment using AH