Eliciting individual risk attitudes – different procedures, different findings

Abstract: We compare three procedures for eliciting individual risk attitudes: Holt-and-Laury (2002), Eckel-and-Grossman (2002), and the general willingness-to-take-risks question of the German socio-economic panel (Dohmen et al. 2011). Using a within-subject design, we carry out a classroom experiment with students who are enrolled in the degree programs Physics, Computer Sciences, Agricultural Sciences, Law, and History. We find that the risk attitudes as measured by the three procedures diverge substantially. This poses a serious challenge to the validity of these measurement instruments. Reconnoitering the room for improvement in the elicitation of risk attitudes, we discuss potential reasons for this divergence such as flawed design, overhasty interpretations, and confusion on the conceptual level.

Keywords: Risk-attitude; Eckel-and-Grossman procedure; Holt-and-Laury procedure; self-reported willingness-to-take-risks; bounded rationality

Contributing authors

Dr. Sven Grüner (Corresponding Author)

Martin Luther University Halle-Wittenberg, Faculty of Natural Sciences III

Institute of Agricultural and Nutritional Sciences, Chair of Agribusiness Management

Karl-Freiherr-von-Fritsch-Str. 4, D-06120 Halle (Saale)

sven.gruener@landw.uni-halle.de

Prof. Dr. Norbert Hirschauer

Martin Luther University Halle-Wittenberg, Faculty of Natural Sciences III

Institute of Agricultural and Nutritional Sciences, Chair of Agribusiness Management

Karl-Freiherr-von-Fritsch-Str. 4, D-06120 Halle (Saale)

norbert.hirschauer@landw.uni-halle.de

Felix Krüger

Martin Luther University Halle-Wittenberg

felix.krueger@student.uni-halle.de

Eliciting individual risk attitudes – different procedures, different findings

Abstract

We compare three procedures for eliciting individual risk attitudes: Holt-and-Laury (2002), Eckel-and-Grossman (2002), and the general willingness-to-take-risks question of the German socio-economic panel (Dohmen et al. 2011). Using a within-subject design, we carry out a classroom experiment with students who are enrolled in the degree programs Physics, Computer Sciences, Agricultural Sciences, Law, and History. We find that the risk attitudes as measured by the three procedures diverge substantially. This poses a serious challenge to the validity of these measurement instruments. Reconnoitering the room for improvement in the elicitation of risk attitudes, we discuss potential reasons for this divergence such as flawed design, overhasty interpretations, and confusion on the conceptual level.

Keywords

Risk-attitude; Eckel-and-Grossman procedure; Holt-and-Laury procedure; self-reported willingness-to-take-risks; bounded rationality

JEL

C91, D81

1 Introduction

In the last decades, economists and other social scientists have developed a wide variety of procedures for eliciting individual risk-attitudes. Despite their diversity, they share the common conceptual background of standard economics. To be more precise, they are based on the assumption of multi-goal decision makers who maximize their expected utility. The expected utility might hinge both on the level and the variability of income. Three types of individuals can be distinguished: *risk-averse* individuals are prepared to trade off income for less variability; in other words, they are prepared to pay a *positive* risk premium. The expected utility of *risk-neutral* individuals depends only on the level of the income irrespective of its variability. Their risk premium is therefore *zero*. *Risk-seeking* individuals enjoy both a high level and a high variability of income. Contrary to risk-averse individuals, they are prepared to trade off income for more variability. They are said to be willing to pay a *negative* risk premium.

Harrison and Rutström (2008) provide a comprehensive survey of procedures for eliciting risk-attitudes used by empirical economists, such as Multiple Price List designs, Random Lottery Pair designs, Ordered Lottery Selection designs, and the Becker-DeGroot-Marschak design. In a *Multiple Price List* design, subjects make a series of choices between gambles that are presented all at once (e.g. Miller et. al. 1969; Holt and Laury 2002). In a *Random Lottery Pair* design, subjects are to make choices between gambles that are presented sequentially (e.g. Hey and Orme 1994; Harrison and

Rutström 2009). In an *Ordered Lottery Selection* design, subjects pick one lottery from an ordered set (e.g. Binswanger 1980, 1981; Eckel and Grossman 2002, 2008). In the *Becker-DeGroot-Marschak* design, subjects are endowed with a lottery for which they have to state a selling price they would accept. If a randomly determined price offer exceeds the subject's selling price, the lottery will be sold; if not, the subject keeps the lottery and plays it out (e.g. Becker, DeGroot, and Marschak 1964; Loomes 1988). Resorting to stated preference approaches, other scholars simply use questionnaires in which subjects are directly asked to self-report their willingness-to-take risks on a pre-specified answer scale (e.g. the risk-attitude question of the German socio-economic panel, SOEP; cf., Dohmen et al. 2011).

Despite the large number and wide variety of procedures for eliciting individual risk attitudes, systematic methodological comparisons have rarely been carried out. One exception is Charness et al. (2013). The authors briefly summarize the evidence found in a range of procedures for eliciting risk attitudes. This systematic review provides interesting insights regarding the differences of various methods. It also indicates that a sound *methodological* comparison requires much more than just comparing studies using different elicitation procedures. One important reason that compromises such comparisons is the fact that these studies vary not only in the elicitation procedure but in many other aspects such as the analyzed social group, the level of monetary incentives, and the context-labels used in the procedures. With this in mind, we make a start by systematically testing and comparing three widespread methods under *ceteris paribus* conditions.

We tackle (1) the Holt-and-Laury (HL) procedure (2002), (2) the Eckel-and-Grossman (EG) procedure (2002), and (3) the self-assessment question as used in the socio-economic panel (SOEP; Dohmen et al. 2011). The incentivized HL and EG procedures can be classified as revealed-preference approaches whereas the self-assessment in the SOEP surveys represents a stated-preference approach. We aim to find out whether these procedures lead to similar results or whether they diverge. Using university students as test persons, we are also interested in finding out whether/how the results of these three procedures vary across different degree programs.

The rest of this paper is structured as follows. We first describe the study design (Section 2) and the subjects and data of our study (Section 3). In Section 4, we present the results of the methodological comparison. In Section 5, we discuss the findings and provide some educated guesses why the three procedures diverge so much. Based on that, we outline some promising lines of future research.

2 Study design

2.1 General study design

We use an experimental *within-subject design* in which each subject is exposed to all three risk-attitude elicitation methods. The potential treatment effects are thus the measured differences in the dependent variable "risk attitude." Compared to a conventional experimental design with random assignment of treatments to subjects (between-subject design), the within-subject design offers both assets and drawbacks. The big advantage of a properly implemented between-subject design is that it eliminates confounding biases – both known and unknown – and thus facilitates internal validity and

causal inference via the experimental design. In other words, a “sufficiently” large experimental population to start with, in conjunction with a random process of treatment assignment, ensures that possible confounders are balanced across treatment groups. This “creates statistical independence between these confounders and treatment assignment” (Dunning 2012: 6). Within-subject designs, in contrast, are potentially threatened by confounding biases that limit causal inference. This is caused by the fact that each subject is sequentially exposed to multiple treatments, which, in turn, may cause order effects and violate the *ceteris paribus* condition (Charness et al. 2012).¹ To mitigate order effects, we randomly determined the sequence in which subjects were exposed to the revealed- and the stated-preference approaches. While one can only attenuate the method’s built-in drawback regarding the *ceteris paribus* condition,² we resort to the within-subject design because of its most important asset, i.e., the time- and cost-saving fact that – compared to a between-subject design – it can more effectively make use of relatively small experimental populations. In our case with 151 evaluable subjects (see section 3), a between-subject design would have provided us with only 50 to 51 subjects in each of three equal-sized randomized treatment groups. In the within-subject design, in contrast, we obtain 151 responses for each of the three treatments.

We recruited students in university classrooms as subjects. We obtained the permission to do so from colleagues teaching major courses in Physics, Computer Sciences, Agricultural Sciences, Law, and History.³ This type of recruitment offers an important advantage: the experimental population represents a group of *pseudo-volunteers* (cf., Eckel and Grossman 2000). While none of the lecturers kept records of attendance, which arguably left us with a “random” sample of students of these courses, none of the students who were present in the classroom refused to participate. This mitigates the self-selection problems that plague other recruitment methods such as the self-selected participation in student subject pools and Internet-based experiments.

2.2 Description of analyzed treatments

Treatment (1): Holt and Laury (HL) procedure

Table 1 illustrates the HL procedure: subjects are confronted with ten subsequent lottery pairs. In each pairing, they have to choose between option A and B. Option A is often referred to as “*safe choice*” because of the small difference between the unfavorable and the favorable payoff (€ 16 vs. € 20). Correspondingly, option B is labeled “*risky choice*” because of the lower payoff of € 1 in the unfavorable case and the higher payoff of € 38.50 in the favorable case. The payoffs remain constant over all lot-

¹ Non-randomized controlled trials based on within-subject designs do therefore not qualify as experiments in a strict sense. They are lacking the essential feature of a true experiment: the randomized assignment of treatments to subjects that facilitates control over confounders *ex ante* via the research design rather than *ex post* via econometric regression models. Rather than introduce new terminology, we nonetheless follow Charness et al. (2012) and others, and retain the label “experimental” for within-subject designs as used in this study.

² Determining the order of exposure for each subject at random does not fully solve the inference problem of within subject designs because – instead of being able to directly compare responses to randomly assigned treatments – one can only compare responses that are averaged over different orderings. For example, the response to SOEP in this study is the aggregate response to orderings where SOEP comes first and where it does not.

³ The idea was to recruit subjects with variation in mathematical skills. We presume that university students of Physics and Computer Sciences are more mathematically versed than students of Law and History.

teries, but the probabilities change systematically. We incentivized the HL procedure to turn it into a revealed-preference approach. Given the low budget that was available, the university students were informed that one student and one lottery pairing⁴ would be randomly chosen and that the lottery chosen by this student would then be played for real.

Table 1: Procedure according to Holt and Laury^(a)

Lottery pairing	Option A: Payoffs (in €) and their probabilities		Option B: Payoffs (in €) and their probabilities		Expected payoff difference (A-B)
1	P(20) = 0.1	P(16) = 0.9	P(38.50) = 0.1	P(1) = 0.9	11.65
2	P(20) = 0.2	P(16) = 0.8	P(38.50) = 0.2	P(1) = 0.8	8.30
3	P(20) = 0.3	P(16) = 0.7	P(38.50) = 0.3	P(1) = 0.7	4.95
4	P(20) = 0.4	P(16) = 0.6	P(38.50) = 0.4	P(1) = 0.6	1.60
5	P(20) = 0.5	P(16) = 0.5	P(38.50) = 0.5	P(1) = 0.5	-1.75
6	P(20) = 0.6	P(16) = 0.4	P(38.50) = 0.6	P(1) = 0.4	-5.10
7	P(20) = 0.7	P(16) = 0.3	P(38.50) = 0.7	P(1) = 0.3	-8.45
8	P(20) = 0.8	P(16) = 0.2	P(38.50) = 0.8	P(1) = 0.2	-11.80
9	P(20) = 0.9	P(16) = 0.1	P(38.50) = 0.9	P(1) = 0.1	-15.15
10	P(20) = 1.0	P(16) = 0.0	P(38.50) = 1.0	P(1) = 0.0	-18.50

(a) The last column was not shown to the subjects.

The more A-choices (*safe choices*) an individual makes the less risk-seeking/more risk-averse an individual is considered to be. Conventionally, 0-3 *safe choices* are seen as reflecting a decreasingly risk-seeking attitude, making 4 *safe choices* is classified as risk-neutral, and 5-9 *safe choices* are understood as reflecting increasing risk-aversion. Equating 4 *safe choices* with risk-neutrality is the conventional interpretation in the literature (e.g. Charness et al. 2013). It should be noted, however, that, in a strict sense, risk-neutrality cannot be deduced from the discrete choices provided in the HL procedure. An individual who makes 4 *safe choices* reveals only that (s)he is to be located somewhere on the range between risk-neutral and slightly risk-averse because changing to the risky choice in the 5th lottery pair already provides this individual with a modest risk premium for the *risky choice*. Since we are interested in understanding how HL – as habitually used *and* interpreted – compares with other methods, we will follow the conventional interpretation and equate 4 *safe choices* with risk-neutrality.

Sometimes, individuals make choices in the HL procedure that are inconsistent with rational choice. For example, picking option A in the 10th “lottery” pair is inconsistent because option B, which offers a certain payoff of € 38.50, strictly dominates option A, which offers a certain payoff of only € 20. We therefore use “lottery” 10 only as “rationality-check device” to exclude inconsistent subjects from the data set. A second inconsistency arises when individuals switch back and forth between lotteries A and B (“multiple switching,” cf., Hirschauer et al. 2016). We drop inconsistently switching subjects from the data set as well.

⁴ Randomly selecting only one task, which is then played out for real, is a common approach in experimental economics and often referred to as *random-lottery incentive system* (e.g. Hey and Lee 2005).

Treatment (2): Eckel and Grossman (EG) procedure

Table 2 illustrates the EG procedure: in contrast to HL, subjects are to choose *one* of several lotteries. Following Dave et al. (2010), we let subjects select one out of six lotteries. From lottery 1 to 6, the unfavorable payoff decreases from € 28 to € 2, and the favorable payoff increases from € 28 to € 70. The probabilities of 50% for the respective lottery's unfavorable and favorable payoffs remain constant across all six lotteries. Analogous to the HL-procedure, students were informed that we would incentivize the EG-procedure, randomly select one student, and play her/his chosen lottery for real.

Table 2: Procedure according to Eckel and Grossman ^(a)

Lottery	Payoffs (in €) and their probabilities		Expected payoff
1	P(28) = 0.5	P(28) = 0.5	28
2	P(24) = 0.5	P(36) = 0.5	30
3	P(20) = 0.5	P(44) = 0.5	32
4	P(16) = 0.5	P(52) = 0.5	34
5	P(12) = 0.5	P(60) = 0.5	36
6	P(2) = 0.5	P(70) = 0.5	36

(a) The last column was not shown to the subjects.

According to conventional interpretation, picking lottery 6 is considered as reflecting risk-seeking behavior.⁵ Choosing lottery 5 is understood as reflecting risk-neutrality, and picking one of the lotteries 4-1 is seen as reflecting increasing risk-aversion. Similar to HL, the conventional interpretation regarding risk-neutrality is a bit too rash. An individual who chooses lottery 5 can only be said to be “slightly risk-averse to risk-neutral.” Again, the discrete differences between the lotteries do not really permit a more clear-cut interpretation because lottery 5 provides the individual with an additional risk premium for the increase in risk compared to lottery 4. Again, we will follow convention and equate the choice of lottery 5 with risk-neutrality because we are interested in the outcome of EG – as habitually used *and* interpreted.

It has been noted that stake sizes affect how people respond to both the HL and the EG procedure (e.g. Holt and Laury 2002; Fehr-Duda et al. 2010). In the HL procedure, the “stake size effect” implies, for example, that an individual who is exposed to lottery pairings with high payoffs such as [€ 160; € 200] vs. [€ 10; € 385] might opt for more safe choices compared to low payoffs of [€ 16; € 20] vs. [€ 1; € 38.5]. Correspondingly, people are assumed to choose lower-numbered lotteries in the EG procedure when stakes are high. To avoid a potentially confounding influence of differing stake sizes within the two revealed-preference approaches, we use comparable payoff magnitudes for both procedures.

Treatment (3): Stated willingness-to-take-risks as used in the SOEP

Table 3 shows the answer scale of the risk-attitude question as used in the SOEP. Subjects are to self-assess their willingness-to-take-risks on an 11-point scale. The wording reads as follows: *How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid tak-*

⁵ Contrary to the quasi-symmetric HL, which distinguishes four different degrees of risk-seeking behavior and five degrees of risk aversion, in EG only one choice is available that denotes a risk-seeking attitude and four choices that describe different degrees of risk aversion. These differing structural designs will cause a divergence between the two methods if people's responses are influenced by a central tendency bias.

ing risks? Please tick a box on the scale, where the value 0 means: “risk averse” and the value 10 means: “fully prepared to take risks.” You can use the values in between to make your estimate.

Table 3: Answer scales of the risk preference question in SOEP

risk-averse											fully prepared to take risks
0	1	2	3	4	5	6	7	8	9	10	
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

According to conventional interpretation, ticking values from 10-6 is considered as denoting a decreasingly risk-seeking attitude, ticking the mid-point 5 is considered as reflecting risk-neutrality, and values from 4-10 are interpreted as increasing risk-aversion. We fear that respondents do often not understand the answer scale as intended by researchers. Two intimately connected reasons might cause confusion: first, the SOEP question opposes the bipolar scales “risk-averse” and “fully prepared to take risks.” Bipolar scales are generally believed to be more difficult for people to understand than unipolar scales that go from “none” to the “extreme.” To avoid ambiguity, unipolar scales are thence to be preferred if possible (cf., e.g. OECD 2013). Second, the very verbalization of endpoints is likely to cause or aggravate misunderstandings. In common language, the label “fully prepared to take risk” is regularly used to describe people who are *nearly* risk-neutral and therefore prepared to take even very large risks even if only a small *positive* risk premium can be expected to be earned. “Fully prepared to take risk” is not usually understood, however, as characterizing risk-seeking behaviors – in the sense used by economists and in the sense used in HL and EG – such as those of gamblers who accept a *negative* risk premium and are prepared to spend money to obtain more risk. Values towards the right-hand side of the SOEP answer scale might therefore represent an inextricable mix of minor risk aversion and truly risk-seeking behaviors. Despite these worries, we will follow the conventional interpretation and take values from 10-6 as denoting a decreasingly risk-seeking attitude because we are interested in comparing SOEP – as habitually used *and* interpreted – with the two other methods.⁶

2.3 Measurement scales of generated data

The numbers produced by HL, EG, and the SOEP risk attitude question are often treated as approximately metric (interval) data, respectively. This is certainly open to criticism, but it seems the rule in many statistical analyses and particularly regressions that use people’s risk attitude either as independent or dependent variable (e.g. Dohmen et al. 2011; Nosić and Weber 2010). In the shaded lines of Table 4, the numbers resulting from HL, EG, and SOEP are systematically arranged – from the highest level of risk seeking (left) to the highest level of risk aversion (right) – to make the basic meaning of these numbers readily comprehensible across the three methods.

⁶ In addition to the general risk-attitude question, subjects were also asked to assess their willingness-to-take-risks in specific domains such as car driving, financial matters, sports and leisure, career, health, and behavior towards unknown people. These “domain-specific” risk attitudes are not directly part of the comparison of the three risk-attitude elicitation procedures, but we will use them when discussing the reasons for their divergence.

Table 4: Numbers produced by HL, EG, and SOEP

		<i>Risk-seeking</i>	<i>Risk-neutral</i>	<i>Risk-averse</i>
1a	HL	0→1→2→3→	4→	5→6→7→8→9
1b		9→8→7→6→	5→	4→3→2→1→0
1c		10→8.75→7.5→6.25→	5→	4→3→2→1→0
2a	EG	6→	5→	4→3→2→1
2b		10→	5→	3.75→2.5→1.25→0
3	SOEP	10→9→8→7→6→	5→	4→3→2→1→0

If one followed the “metric-data argument,” one might be tempted to try to make the numbers from the three methods *directly* comparable through some kind of data transformation. Using the SOEP scale as base line, one would probably first reverse the HL numbers from 0→1→2→3→4→5→6→7→8→9 (see Table 4, line 1a) into 9→8→7→6→5→4→3→2→1→0 (see line 1b). SOEP distinguishes 11 discrete risk attitudes but HL only 10. If one additionally assumed that they cover exactly the same range, one could “stretch” the reversed HL numbers into the series **10→8.75→7.5→6.25→5→4→3→2→1→0** (see line 1c). An analogous exercise would transform the EG series from the original 6→5→4→3→2→1 (see line 2a) into 10→5→3.75→2.5→1.25→0 (see line 2b). We believe, however, that such transformations are not permissible and that they would be grossly misleading. To start with, the numbers produced by the three methods should be treated as ordinal data within each method. But even if we considered the numbers *within* each method as approximately metric, there is simply no way of knowing whether the respective series of numbers really cover identical ranges. There is no way of knowing, for example, whether the original EG number of 6 should be transformed into 6.25 or 10 or some other number between 6 and 10 to make it equivalent to the numbers in SOEP.

In our examination of whether HL, EG, and SOEP lead to similar results or whether they diverge, we therefore cannot help but treat the data produced by these methods as nominal data and use only the three categories “risk-seeking,” “risk-neutral,” and “risk-averse” in the analysis.

3 Description of subjects

Our experimental study was carried out as pen-and-paper exercise in class from late November to mid-December 2017 at the Martin Luther University Halle-Wittenberg (Germany).⁷ A total of 170 students in 5 classrooms were recruited as experimental subjects. We dropped 19 students from the analysis because of inconsistent answering.⁸ The descriptive statistics of the remaining 151 subjects is shown in Table 5.

⁷ To reduce the risk of confounding through an experimenter effect, the study was carried out by the same experimenter in all five classrooms.

⁸ A variant calculation showed that our key findings remain robust irrespective of whether we keep inconsistent subjects in the analysis or not.

Table 5: Description of subjects ^(a)

Degree program	Physics (N=26)	Computer Sciences (N=25)	Agricultural Sciences (N=31)	Law (N=36)	History (N=21)	Others ^(b) (N=15)	Total ^(c) (N=151)
Share of male subjects (in %)	57.7	80.0	51.6	47.2	52.4	80.0	59.0
Age (in years)	20.0	21.2	21.7	19.0	23.2	22.5	20.9
Semester of study	3.4	3.8	4.0	1.2	4.5	4.8	3.3
Raised in rural areas (in %)	40.0	54.2	77.0	47.2	71.4	60.0	58.1
Life satisfaction ^(d)	6.9	7.2	7.3	7.4	7.0	7.2	7.2
Current mood ^(e)	3.5	4.0	3.9	3.7	3.6	4.0	3.8

(a) Not all subjects (students) answered all questions.

(b) Students of Mathematics (5), Philosophy (1), Economics (1), Political Sciences (6), and Sociology (2) who also attended one the five major courses are collectively referred to as “Others.”

(c) The number of students indicated in the degree programs exceeds the total of 151 because of double-enrolments.

(d) “Life satisfaction” was measured on an 11-point scale from 0 (=not at all satisfied) to 10 (=completely satisfied).

(e) “Current mood” was measured on a 5-point scale from 1 (=very poor) to 5 (=very good).

On average over all degree programs, 59% of students were male. The share of male students was higher in the mathematically oriented degree programs Physics (57.7%) and Computer Sciences (80.0%) than in the non-mathematical degree programs Law (47.2%) and History (52.4%). The subjects were on average approximately 21 years old and in the 4th semester of their respective degree program. Slightly more than 50% of the students were raised in rural areas; as might have been expected, the fraction was particularly high with students enrolled in Agricultural Sciences (77%). Somewhat surprisingly, the fraction was also quite high with students of History (71.4%). On an 11-point scale ranging from 0 (=not at all satisfied) to 10 (=completely satisfied), the students self-reported an average life satisfaction of 7.2. Self-reported life satisfaction was lowest among students of Physics (6.9) and highest among students of Law (7.4), with the students in other programs being between 7.0 and 7.3. A similar pattern was observed for the current mood of students. On a 5-point scale ranging from 1 (=very poor) to 5 (=very good), students reported a “relatively good mood” (3.8) on average. Again, students of Physics scored lowest across the different degree programs.

4 Data analysis

4.1 Cross-comparisons of the three risk-attitude elicitation procedures

To get a quick picture of how much HL, EG, and SOEP diverge, we first take a look at the share of subjects that these procedures classify as risk-averse, risk-neutral, and risk-seeking (cf., Table 6). While all procedures find the majority of individuals to be risk-averse, the levels vary considerably: from 52.4% in SOEP over 61.6% in HL to 79.5% in EG. Similarly, large deviations between the three procedures are found for the share of subjects who are classified as risk-neutral and risk-seeking.

Table 6: Share of subjects classified as risk-averse, risk-neutral, and risk-seeking according to HL, EG, and SOEP

	HL (N=151) ^(a)	EG (N=151) ^(a)	SOEP (N=145) ^(a)	Weighted average ^(b)
Risk-averse subjects (in % of sample) (Deviation in % of mean)	61.6 (-5)	79.5 (+23)	52.4 (-19)	64.7
Risk-neutral subjects (in % of sample) (Deviation in % of mean)	17.9 (+18)	12.6 (-17)	15.2 (0)	15.2
Risk-seeking subjects (in % of sample) (Deviation in % of mean)	20.5 (+2)	8.0 (-60)	32.4 (+61)	20.1

(a) The number of subjects differ between the procedures because not all subjects gave answers in all of them.

(b) The weighted average is based on the number of subjects who responded to the respective procedures.

Looking across the categorizations produced by the three procedures, one cannot help but state disaffectedly that they diverge *substantially* and that there seems to be not much of an identifiable pattern regarding this divergence. Despite their widespread use in studies tackling the influence of people's risk attitudes on decisions under risk, this outcome substantially challenges the assumption that these procedures represent adequate methods for the elicitation of people's individual risk attitudes. This holds even though we necessarily content ourselves with a comparison of nominal categorizations and thus the lowest level of measurement.

There are two exceptions where there seems to be a pattern of divergence that is understandable in the light of their inherent methodical weaknesses as discussed in Section 2: first, according to SOEP, which is based on the self-reported willingness-to-take-risks, far less subjects would be classified as risk-averse and far more subjects as risk-seeking compared to the two lottery-based procedures HL and EG. This can be taken as an indication that an unknown but substantial fraction of subjects did not understand the label “fully prepared to take risk” as describing risk-seeking in the sense used by economists, but rather as being prepared to take even very large risks even if only very small risk premiums can be earned. The group of subjects who would be classified as risk-seeking according to the conventional interpretation of SOEP might therefore indeed represent an inextricable mix of individuals who are “weakly risk-averse” or indeed risk-seeking. Second, according to EG, particularly few individuals are classified as risk-seeking and particularly many as risk-averse compared to HL and SOEP. This might be caused by its asymmetric architecture that provides only one choice for risk-seeking individuals but four choices for risk-averse individuals (see footnote 5). While we might assume that all three methods are likely to suffer to a certain degree from central tendency bias (cf., Choi and Pak 2005), its effects will differ markedly. It will cause too many individuals to be classified as risk-neutral in the more or less symmetric HL and SOEP designs. In the EG design, however, it will cause too many individuals to be classified as risk-averse.

To obtain more details regarding the divergence of the three elicitation methods, we additionally make a pairwise comparison for each of three risk attitude categories. Table 7a compares HL with SOEP, Table 7b contrasts HL with EG, and Table 7c compares EG with SOEP.

Table 7a: Cross-comparisons of the procedures: HL/SOEP (N=145) ^(a)

		SOEP		
		Risk-averse	Risk-neutral	Risk-seeking
HL	Risk-averse	39.3%	7.6%	13.1%
	Risk-neutral	8.3%	3.5%	16.9%
	Risk-seeking	4.8%	4.1%	12.4%

Table 7b: Cross-comparisons of the procedures: HL/EG (N=151) ^(a)

		EG		
		Risk-averse	Risk-neutral	Risk-seeking
HL	Risk-averse	52.3%	8.0%	1.3%
	Risk-neutral	13.9%	0.7%	3.3%
	Risk-seeking	13.3%	4.0%	3.3%

Table 7c: Cross-comparisons of the procedures: EG/SOEP (N=145) ^(a)

		SOEP		
		Risk-averse	Risk-neutral	Risk-seeking
EG	Risk-averse	46.2%	10.3%	22.1%
	Risk-neutral	4.8%	2.8%	5.5%
	Risk-seeking	1.4%	2.1%	4.8%

(a) The number of subjects differ between the comparisons because not all subjects gave answers in all procedures.

The following information can be derived from these three contingency tables:

1. The numbers in the gray-shaded fields indicate the degree to which the methods conform with each other in that they categorize subjects identically. For example, only 39.3% of subjects are categorized as risk-averse according to both the HL and the SOEP procedure (Table 7a). The respective numbers are 52.3% and 46.2% in the HL/EG and EG/SOEP comparison (Table 7b and Table 7c).
2. If two procedures conformed completely with each other, the sum of the percentage values in the gray-shaded fields from top left to bottom right would add up to 100%, and all other fields would be empty. This is why we use the “diagonal percentage sum” as an overall measure of pairwise conformity. Pairwise conformity amounts to 55.2%, 56.3%, and 53.8% in the HL/SOEP, HL/EG, and EG/SOEP comparison, respectively. In other words, pairwise conformity is rather poor, but it does not differ very much across the three comparisons.⁹

⁹ When describing the elicitation procedures in 2.2, we pointed out that risk-neutrality, which describes *one* discrete value of the continuous state variable “risk attitude,” cannot be deduced, in a strict sense, from the discrete choices provided in the HL and EG procedure. Taking this seriously, would imply to merge the categories “risk-averse” and “risk-neutral” in HL and EG. While representing much coarser information, using only two

3. While the non-conformity level is similar in all three comparisons, it results from different types of divergence. Let us look at antipodal classifications only. In the HL/SOEP comparison, 17.9% of subjects (=13.1% + 4.8%) are classified as *risk-averse* according to one method but as *risk-seeking* according to the other (Table 7a). The respective numbers are 14.6% and 23.5% in the HL/EG and EG/SOEP comparison (Table 7b and Table 7c). With 22.1%, the highest share of antipodal classification are subjects who are classified as risk-averse in EG but as risk-seeking in SOEP. This can be related to the pattern of divergence that we already discussed above: first, too many individuals are likely to be classified as risk-seeking in SOEP because an unknown fraction of subjects understands the label “fully prepared to take risk” as “very little risk-averse.” Second, due to its asymmetric architecture, too many individuals are likely to be classified as risk-averse in EG if subjects’ responses – as might be expected – are affected by a central tendency bias.

4.2 Risk attitude elicitation across degree programs and elicitation procedures

In this section, we focus on how the elicited risk attitudes vary across the three elicitation procedures and the five degree programs (classrooms) from which subjects were recruited.

Perceived complexity of the EG and HL procedures

In their survey, Charness et al. (2013) distinguish simple and complex procedures for eliciting individual risk attitudes. They classify HL as “complex” and EG as “simple.” Since complexity and thus comprehensibility might affect subjects’ responses to these procedures, we asked the participating students how accessible they found HL and EG. Based on a 5-point scale ranging from very difficult to understand (=1) to very easy to understand (=5), Table 8 describes the average complexity of the HL and EG procedures as perceived by the students in the different degree programs.

Table 8: Perceived complexity of HL and EG ^(a)

		Physics (N=28)	Computer Sciences (N=26)	Agricultural Sciences (N=33)	Law (N=40)	History (N=29)	Total ^(b) (N=155)
Perceived complexity EG	Mean (SD)	4.5 (0.7)	4.7 (0.6)	4.7 (0.6)	4.3 (0.8)	4.1 (0.9)	4.5 (0.7)
Perceived complexity HL	Mean (SD)	4.1 (0.8)	4.3 (0.7)	4.0 (1.1)	3.7 (0.9)	3.5 (1.1)	3.9 (1.0)
Inconsistent subjects in HL		7.1%	3.9%	6.1%	10.0%	27.6%	11.0%

(a) Perceived complexity was measured on a 5-point scale ranging from very difficult (=1) to very easy to understand (=5).

(b) From the total of 170 students, 15 students were excluded because they were enrolled in other degree programs.

(c) The sum of students indicated in the degree programs exceeds the total of 138 by one because of a double-enrolment.

Across all five degree programs, students found EG easier to understand than HL. Students of the mathematically oriented programs (Physics and Computer Sciences), found the EG and particularly the HL procedure more accessible than students of the non-mathematical majors (Law and History). In

categories in HL and EG would increase conformity from 55.2% to 71.0% in the HL/SOEP comparison, from 56.3% to 78.1% in the HL/EG comparison, and from 53.8% to 69.0% in the EG/SOEP comparison.

line with that, students of Law and especially History made more inconsistent choices in the HL procedure than students of Physics and Computer Sciences.

Classification of students in different programs according to the three elicitation procedures

Do the results of the three procedures vary across different degree programs and, if so, to what extent? Trying to answer these questions, we depict the share of students in the different majors who are classified as risk-averse, risk-neutral, and risk-seeking according to HL, EG, and SOEP (cf., Table 9).

Table 9: Share of individuals classified as risk-averse, risk-neutral, and risk-seeking according to HL, EG, and SOEP (in %)

		Physics (N=26)	Computer Sciences (N=25)	Agricultural Sciences (N=31) ¹	Law (N=36)	History (N=21) ^(a)	Total ^(a) (N=138)
HL	risk-averse	73.1	68.0	52.8	52.8	47.6	60.2
	risk-neutral	11.5	24.0	19.4	19.4	19.1	18.8
	risk-seeking	15.4	8.0	27.8	27.8	33.3	21.0
EG	risk-averse	88.5	64.0	77.8	77.8	81.0	78.3
	risk-neutral	7.7	24.0	16.7	16.7	9.5	13.8
	risk-seeking	3.9	12.0	5.6	5.6	9.5	7.9
SOEP ^(b)	risk-averse	60.0	45.8	51.5	51.5	42.9	52.3
	risk-neutral	8.0	25.0	12.1	12.1	23.8	15.2
	risk-seeking	32.0	29.2	36.7	36.7	33.3	32.5
Mean ^(c)	risk-averse	73.9	59.3	60.7	60.7	57.2	63.8
	risk-neutral	9.1	24.3	16.1	16.1	17.5	15.9
	risk-seeking	17.1	16.4	23.4	23.4	25.4	20.3

(a) From the total of 170, we dropped 19 students due to inconsistent answers. Due to their low sample sizes, we excluded 15 students from the degree programs Mathematics, Philosophy, Economics, Political Sciences, and Sociology, who were also present in the five classrooms. Two students both provided inconsistent answers and were enrolled in other programs. Furthermore, the sum of students indicated in the five degree programs exceeds the total of 138 by one because of a double-enrolment.

(b) The numbers in the column headings apply for HL and EG. The student numbers answering the SOEP question are as follows: Physics (25), Computer Sciences (24), Agricultural Sciences (30), Law (33), and History (21).

(c) Weighted average based on the number of students who responded to the respective procedures.

There is no eye-catching pattern regarding the risk-attitude categorizations produced by the three elicitation procedures across the five degree programs. At first view, the differentiated view across the programs simply seems to reconfirm two basic findings already known from the summary Table 6: first, the majority of students is categorized as risk-averse across all three procedures; and this holds for all degree programs. Second, the share of the three categories “risk-averse,” “risk-neutral,” and “risk-seeking” varies considerably from one elicitation procedure to the other; and this holds also for all degree programs. Nonetheless, some specific findings regarding the degree programs can be identified:

1. According to all three procedures, the largest share of risk-averse students and the lowest share of risk-neutral ones are found among students of Physics. Again, there is considerable variation

across the three procedures. For example, 73.1% of Physics students are classified as risk-averse according to HL, 88.5% according to EG, and 60.0% according to SOEP. Comparable variations across the procedures can be found in the other degree programs.

2. The share of students classified as risk-averse according to HL is higher in the mathematically oriented degree programs Physics and Computer Sciences compared to the non-mathematical degree programs Law and History. The opposite holds for the risk-seeking category. Considerably less students of Physics and Computer Sciences are classified as risk-seeking compared to the students of Law and History. However, these patterns cannot be found in the other procedures.
3. All procedures classify 24-25% of the students of Computer Sciences as risk-neutral. This is the highest share of risk-neutral students among all degree programs.

In a nutshell, the differentiated view across the programs reveals no clear pattern. While the risk attitude categorizations that are attributed to the students by the three procedures vary across the different degree programs, program-specific findings remain tentative and resist an easy and clear-cut explanation.

5 Discussion and outlook

Examining the Holt-and-Laury procedure (HL), the Eckel-and-Grossman procedure (EG), and the self-reported willingness-to-take-risks question used in the socio-economic panel (SOEP), we find substantial differences in the risk attitude categorizations that are attributed by these procedures to the test persons (students) whom we recruited from various university degree programs. While our study was small and while we looked at a convenience sample of students, this divergence poses a substantial challenge for the all-too-rarely questioned assumption that these widely used procedures represent adequate measurement instruments for people's individual risk attitudes.

The large divergence between the three procedures may be caused by various reasons. Even a cursory look at the procedures (measurement instruments) suggests that their design might be flawed in that subjects do not fully understand their tasks or the question as intended by researchers. This is reflected in our findings: especially students from non-mathematical majors such as Law and History found the quantitative and money-related choices in EG and, in particular, in HL difficult to understand. They also made more inconsistent choices than students of Physics or Computer Sciences. We must realize that, if subjects do not grasp the tasks in HL and EG, we measure "something" (which we do not know) but not the theoretical construct "risk attitude" we are interested in. Poor understanding is likely to be more pronounced in the "complex" HL because serial choices have to be made under varying outcome *probabilities* – a concept whose meaning is believed to be unclear to many people (Gigerenzer 2015).

Despite its seeming simplicity compared to HL and EG, the SOEP willingness-to-take-risks question may be threatened by even more misunderstandings. We fear that respondents do often not understand the bipolar answer scale from "risk averse" to "fully prepared to take risks as intended by researchers. Economists interpret "fully prepared to take risks" as risk-seeking behavior in terms of being prepared

to spend money to obtain more risk (negative risk premium). In common language, however, this wording is usually used for people who are nearly risk-neutral and therefore prepared to take large risks even if only a small positive risk premium can be earned. Answers on the SOEP answer scale that are conventionally interpreted as “risk-seeking” are therefore likely to be an inextricable mix of minor risk aversion and truly risk-seeking behavior, which, in turn, can cause a divergence compared to the results of the other procedures.

As in other fields of research based on people ticking answer boxes, responses to all three risk-attitude elicitation procedures are likely to be affected by a central tendency bias. However, the effects of a central tendency bias will differ markedly between the procedures. In the asymmetric design of EG, too many individuals will be classified as risk-averse. In the quasi symmetric HL and SOEP designs, too many individuals will be classified as risk-neutral. A central tendency bias will therefore cause divergence between the procedures.

Besides misunderstandings by respondents, the validity of the measurement instruments HL and EG might be threatened through an overhasty interpretation. Conventionally, 4 safe choices in HL and choosing lottery 5 in EG are interpreted as risk-neutrality. However, risk-neutrality, which describes *one* discrete value of the continuous state variable “risk attitude,” cannot really be deduced from the discrete variations provided in HL and EG. Instead, the said two choices already entail a positive albeit small risk premium. Individuals who make these choices can only be said to be “slightly risk-averse to risk-neutral.” In a strict sense, researchers using HL and EG should therefore content themselves with the coarser information contained in the two categories “risk-averse” and “risk-seeking”.

Finally, beyond misunderstandings and misinterpretations, a conceptual confusion may contribute to the divergence between the procedures that, each by itself, presumably elicits *the* risk attitude of an individual. Contradictory to the concept of a unique individual risk attitude, the students’ answers to the context-specific willingness-to-take-risks question in SOEP suggest that people might exhibit different risk attitudes in different life domains. A more pronounced willingness-to-take-risks was indicated by respondents in the sports and leisure domain compared to financial matters. This raises the question of what people have in mind when they respond to the general (i.e., non-domain specific) willingness-to-take-risks question. Some may think about activities that are particularly relevant for them at present; others may simply associate the activity they had been engaged in shortly before answering the question (availability bias). As a result, respondents’ answers might be a mix of different domain-specific risk attitudes that the researcher is unable to distinguish. This will contribute to the divergence of the risk attitudes as “found” by the respective procedure if respondents typically associate different domains with the different framings.

Replications – ideally with larger numbers of subjects and ideally with representative samples of various social groups of interest – are necessary to find out whether our results, which pose a serious challenge to the validity of risk attitude elicitation procedures, are robust or not. If the large divergence in the results of different elicitation procedures is found to be robust in further studies, researchers should take this as a cue that they need to carefully select procedures or even think about the development of

modified/new procedures that represent an adequate fit to the respective object of investigation. There are two aspects that seem to require particular attention: first, the validity of quantitative procedures such as HL is endangered if subjects are not well-versed in mathematics and probability theory. Such quantitative methods should therefore only be used if a well justified assumption can be made that the individuals under investigation do understand the concept of probability. In the case of non-mathematically versed subjects, they should be modified in that didactical explanations are provided to respondents that bring about a proper understanding of the tasks. Second, it is important to critically scrutinize whether people mentally associate specific life domains with the framings of the respective elicitation procedures. To avoid answers that represent a conglomerate of different domain-specific risk attitudes, researchers should consider to substitute general or “context-free” framings by framings that explicitly describe the risky decision context for which the risk attitude is to be elicited. Both aspects may be amplified if the subjects under investigation are not university students. For example, professionals can often be expected to be more heterogeneous regarding their skills and spontaneous associations than students. Providing for a proper understanding and ensuring an association with the life domain envisaged by the researcher becomes therefore both more difficult and more important when trying to elicit the risk attitudes of individuals who are not students.

Acknowledgment

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 388911356.

References

- Becker, G.M., DeGroot, M.H., Marschak, J. (1964): Measuring utility by a single-response sequential method. *Behavioral Science* 9: 226–232.
- Binswanger, H.P. (1980): Attitudes toward risk: Experimental measurement in rural India. *American Journal of Agricultural Economics* 62: 395–407.
- Binswanger, H.P. (1981): Attitudes toward risk: Theoretical implications of an experiment in rural India. *Economic Journal* 91: 867–890.
- Charness, G., Gneezy, U., Imas, A. (2013): Experimental Methods: Eliciting Risk Preferences. *Journal of Economic Behavior & Organization* 87: 43–51.
- Charness, G., Gneezy, U., Kuhn, M.A. (2012): Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization* 81(1): 1–8.
- Choi, B.C.K., Pak, A.W.P. (2005): A Catalog of Biases in Questionnaires. *Preventing Chronic Disease* 2(1): A13.
- Dave, C., Eckel, C.C., Johnson, C.A., Rojas, C. (2010): Eliciting risk preferences: when is simple better? *Journal of Risk and Uncertainty* 41(3): 219–243.
- Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., Wagner, G.G. (2011): Individual risk attitudes: measurement, determinants and behavioral consequences. *Journal of the European Economic Association* 9: 522–550.
- Dunning, T. (2012): *Natural Experiments in the Social Sciences: A Design-based Approach*. Cambridge, Cambridge University Press.

- Eckel, C.C., Grossman, P.J. (2000): Volunteers and pseudo-volunteers: The effect of recruitment method in dictator experiments. *Experimental Economics* 3(2): 107–120.
- Eckel, C.C., Grossman, P.J. (2002): Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior* 23(4): 281–295.
- Eckel, C.C., Grossman, P.J. (2008): Forecasting risk attitudes: An experimental study of actual and forecast risk attitudes of women and men. *Journal of Economic Behavior & Organization* 68(1): 1–17.
- Fehr-Duda, H., Bruhin, A., Epper, T., Schubert, R. (2010): Rationality on the rise: Why relative risk aversion increases with stake size. *Journal of Risk and Uncertainty* 40(2): 147–180.
- Gigerenzer, G. (2015): *Risk Savvy: How to Make Good Decisions*. New York, Penguin Books.
- Harrison, G.W., Rutström, E.E. (2005): Expected utility theory and prospect theory: one wedding and a decent funeral. *Experimental Economics* 12: 133–158.
- Harrison, G.W., Rutström, E.E. (2008): Risk Aversion in the Laboratory. In: Cox, J.C., Harrison, G.W. (eds.): *Risk Aversion in Experiments (Research in Experimental Economics, Volume 12)*. Bingley, Emerald Group Publishing Limited: 41–196.
- Hey, J.D., Lee, J. (2005): Do Subjects Separate (or Are They Sophisticated)? *Experimental Economics* 8(3): 233–265.
- Hey, J.D., Orme, C. (1994): Investigating generalizations of expected utility theory using experimental data. *Econometrica* 62(6): 1291–1326.
- Hirschauer, N., Musshoff, O., Maart-Noelck, S.C., Gruener, S. (2014): Eliciting risk attitudes – how to avoid mean and variance bias in Holt-and-Laury lotteries. *Applied Economics Letters* 21(1): 35–38.
- Holt, C.A., Laury, S.K. (2002): Risk aversion and incentive effects. *American Economic Review* 92(5): 1644–1655.
- Loomes, G. (1988): Different experimental procedures for obtaining valuations of risky actions: Implications for utility theory. *Theory and Decision* 25: 1–23.
- Miller, L., Meyer, D.E., Lanzetta, J.T. (1969): Choice among equal expected value alternatives: Sequential effects of winning probability level on risk preferences. *Journal of Experimental Psychology* 79(3): 419–423.
- Nosić, A., Weber, M. (2010): How Riskily Do I Invest? The Role of Risk Attitudes, Risk Perception, and Overconfidence. *Decision Analysis* 7(3): 282–301.
- OECD (2013): *OECD Guidelines on Measuring Subjective Well-being*. Paris, OECD Publishing. <https://doi.org/10.1787/9789264191655-en>.
- Tversky, A., Kahneman, D. (1981): The framing of decisions and the psychology of choice. *Science* 211: 453–458.
- Vogt, W.P., Vogt, E.R., Gardner, D.C., Haefele, L.M. (2014): *Selecting the right analyses for your data: quantitative, qualitative, and mixed methods*. New York, The Guilford Publishing.