ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Holzmeister, Felix; Stefan, Matthias

Working Paper

The risk elicitation puzzle revisited: Across-methods (in)consistency?

Working Papers in Economics and Statistics, No. 2019-19

Provided in Cooperation with:
Institute of Public Finance, University of Innsbruck

Suggested Citation: Holzmeister, Felix; Stefan, Matthias (2019) : The risk elicitation puzzle revisited: Across-methods (in)consistency?, Working Papers in Economics and Statistics, No. 2019-19, University of Innsbruck, Research Platform Empirical and Experimental Economics (eeecon), Innsbruck

This Version is available at:
https://hdl.handle.net/10419/207084

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

universität innsbruck

Research Platform
Empirical and Experimental Economics

working paper

eeecon
[triple:e:con]

The risk elicitation puzzle revisited: Across-methods (in)consistency?

Felix Holzmeister, Matthias Stefan

Working Papers in Economics and Statistics

2019-19

University of Innsbruck
https://www.uibk.ac.at/eeecon/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Banking and Finance
- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact address of the editor:
research platform “Empirical and Experimental Economics”
University of Innsbruck
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 71022
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
https://www.uibk.ac.at/eeecon/wopec/

For a list of recent papers see the backpages of this paper.

The risk elicitation puzzle revisited: Across-methods (in)consistency?

Felix Holzmeister[†,*] Matthias Stefan[†]

[†] Department of Banking and Finance, University of Innsbruck
[*] Corresponding author: felix.holzmeister@uibk.ac.at

Abstract

With the rise of experimental research in the social sciences, numerous methods to elicit and classify people's risk attitudes in the laboratory have evolved. However, evidence suggests that people's attitudes towards risk may change considerably when measured with different methods. Based on a with-subject experimental design using four widespread risk preference elicitation methods, we find that different procedures indeed give rise to considerably varying estimates of individual and aggregate level risk preferences. Conducting simulation exercises to obtain benchmarks for subjects' behavior, we find that the observed heterogeneity in risk preference estimates across methods looks qualitatively similar to the heterogeneity arising from independent random draws from choices in the experimental tasks, despite significantly positive correlations between tasks. Our study, however, provides evidence that subjects are surprisingly well aware of the variation in the riskiness of their choices. We argue that this calls into question the common interpretation of variation in revealed risk preferences as being inconsistent.

JEL: C91, D81
Keywords: Risk preference elicitation, inconsistent behavior, risk attitudes

We thank Antonio Filippin, Christoph Huber, Jürgen Huber, Michael Kirchler, Michael Razen, Julia Rose, Matthias Sutter, Roberto Weber, Erik Wengström, and two anonymous referees, participants at the research seminar at the Max Planck Institute in Bonn, the Experimental Finance Conference 2018 in Heidelberg, the Economic Science Association Conference 2018 in Berlin, the Conference on Decision Sciences 2018 in Konstanz, and the Nordic Conference on Behavioral and Experimental Economics 2018 for helpful comments and suggestions to improve the manuscript. Financial support from the Austrian Science Fund FWF (SFB F63), and the University of Innsbruck (Aktion D. Swarovski KG) is gratefully acknowledged.

"You are—face it—a bunch of emotions, prejudices, and twitches,
and this is all very well as long as you know it."

—Adam Smith (1968), *The Money Game.*

1. Introduction

Risk is an integral part of many economic decisions and, thus, has been considered a key building block of economic theory (Arrow, 1965). As a consequence, the question how to properly elicit and classify individuals' risk preferences is of vital importance in academic research. In experimental economics and psychology, irrespective of differences in their approach, incentiviced risk preference elicitation tasks have evolved as widely accepted tools to measure and assess individual-level attitudes towards risk. While economists and psychologists have developed a variety of competing methodologies, a consensus on which of the elicitation procedures gives rise to the most accurate estimates of individual-level risk preferences has not been reached yet (Charness et al., 2013). Facing this pluralism of methods, pragmatism prevails among researchers when choosing among various competing methods. The implicit assumption behind this common practice is the procedural invariance axiom, which states that normatively equivalent elicitation methods should give rise to the same preference ordering (Tversky et al., 1988). The experimenter's choice of which method to use should thus not systematically affect participants' revealed risk preferences. However, experimental evidence, reviewed in detail in Section 2, suggests that participants' attitudes towards risk may vary considerably when measured with different elicitation methods—a finding recently referred to as the "risk elicitation puzzle" (Pedroni et al., 2017).

Given the abundance of experimental findings on variations of risk preferences across methods, the following question arises: How can we assess whether subjects would actually want to choose invariably across task but are driven towards inconsistent decisions by external factors, or, whether different methods rather stimulate diverging preference revelations? In other words, how can we assess whether we are facing inconsistent behavior or rather varying (sets of) preferences of experimental subjects? The path taken in this paper takes into account the subjective point of view of the experimental participants: In addition to incentivized risk preference elicitation methods, our experimental protocol comprises survey items, which allow for examining whether participants are aware of the risk levels they take in the different tasks.

In particular, in our experimental study we use a within-subject design incorporating four widely used risk preference elicitation methods: (i) the "bomb" risk elicitation task (Crosetto and Filippin, 2013), the certainty equivalent method (Abdellaoui et al., 2011; Cohen et al., 1987; Dohmen et al., 2010), a multiple choice list between pairs of lotteries (Holt and Laury, 2002, 2005), and a single choice list (Binswanger, 1980, 1981; Eckel and Grossman, 2002, 2008). In order to obtain a more encompassing picture of across-methods variation in risk preferences, we study a set of intra- and inter-subject measures. We observe that subjects' revealed preferences are consistent in less than 50% of pairwise comparisons of methods. Conducting simulation exercises to obtain benchmarks for subjects' behavior, we find that the observed heterogeneity of risk preference estimates across methods looks qualitatively similar to the heterogeneity arising from independent random draws from choices in the experimental tasks, despite significantly positive correlations in risk-taking between the tasks. While we report evidence for

substantial across-methods variation in risk-taking behavior, our main result is that subjects' assessments of the riskiness of their choices is significantly related to the risk preference estimates across the different tasks. Thus, subjects seem to be well aware of their seemingly inconsistent choices across methods. In the light of these results, we argue that the observed variation in revealed preferences cannot be straightforwardly interpreted as being inconsistent.

Despite being a common assumption in experimental economics and psychology, the procedural invariance axiom is challenged by a vast number of findings reporting variations in revealed risk preferences across different elicitation methods. Yet, it is not *per se* clear where these variations stem from: (i) The decision environment implied by different risk preference elicitation methods might stimulate different preferences in subjects. If this should be the case, the procedural invariance axiom is directly challenged; and indeed, calling procedural invariance into question dates back to early systematic examinations of preference reversals (see e.g. Tversky et al., 1988; Tversky and Thaler, 1990). (ii) If a subject's choices in two tasks can be described by different risk preferences, her behavior might be inconsistent—a term abundantly used in the literature. However, it is not immediately obvious what the term *inconsistent* should refer to in terms of choice behavior. As argued by Sen (1993), "the basic difficulty arises from the implicit presumption underlying that approach that *acts* of choices are, on their own, like statements which can contradict, or be consistent with, each other." To assess consistency of behavior, eventually, one needs to invoke a theory upon which choices can be interpreted as contradictory (Sugden, 1991). This is an essential insight, not least regarding the risk elicitation puzzle, as it illustrates that one can only assess the consistency of choices across different methods on the basis of some underlying theoretical framework. Part of this theoretical framework is the procedural invariance axiom, which allows for evaluating diverging behavior as inconsistent under the assumption that different methods should elicit the same preferences. If this assumption is omitted, i.e., if we suppose that different preference relations dictate choices across different methods, the classification of inconsistent behavior becomes obsolete. The results reported in our paper should serve as an invitation to reconsider and reassess the assumption of procedural invariance of methods.

2. Related Literature

Following the existing literature, in this section we refer to behavior revealing varying risk preferences as "inconsistent" without thereby adopting an interpretation of the observed behavior.[1]

The question whether different experimental procedures to measure individual risk attitudes give rise to the same revealed preferences dates back more than 50 years. Slovic (1964), to the best of our knowledge, was first to challenge the standard assumption of procedural invariance by concluding that *"the domain of risk taking behavior may not be as conceptually unitary as many psychologists would like to believe."* An early study by Slovic (1972a) comparing attitudes towards risk using two different procedures corroborates the skepticism about method invariance by emphasizing low levels of inter-measure

[1] Please note that our outline of the related literature comprises results from the economic and the psychological literature alike. While the two fields may differ in their methodological approach, e.g., regarding the focus on normative aspects of preference elicitation or the external validity of different measures, we deem these distinctions of secondary importance for a summary of the evidence on (seemingly) inconsistent behavior in incentivized risk preference elicitation methods.

correlation. Slovic (1972a,b) argues that different procedures trigger different processing of information about probabilities and payoffs and that situation specificity is a crucial dimension of risk-taking behavior.

Almost three decades later, the question whether risk preferences are properly modelled as a generally stable personality trait has been revisited. Using a first price auction and a Becker-DeGroot-Marschak procedure (BDM; Becker et al., 1964), Isaac and James (2000) find that the rank-order of revealed preferences across individuals is not preserved across the two institutions. Berg et al. (2005) substantiate these results in a non-parametric framework, comparing revealed risk preferences in a BDM-mechanism, an English clock auction, and a first price auction. In a similar manner, several more recent studies investigate across-methods consistency of risk preferences utilizing multiple price list formats. Anderson and Mellor (2009) show that subjects do not reveal stable risk preferences across an incentivized price list (HL; Holt and Laury, 2002) and an unincentivized survey on hypothetical gambles. Bruner (2009) reports pronounced inconsistencies in choices between two price lists with the same expected payoffs, only altering whether lotteries vary in payoff or probability. Hey et al. (2009) examine stability of revealed preferences across four different elicitation methods and conclude that the differences in the methods' noisiness and bias might account for inconsistencies. Dave et al. (2010) and Reynaud and Couture (2012) compare risk preferences estimated with the HL-method and the single choice list procedure introduced by Eckel and Grossman (2002). Both studies report substantial differences in estimated risk coefficients. While Dave et al. (2010) suggest that inter-subject differences in risk preference estimates can partly be attributed to a lack of numeracy, Reynaud and Couture (2012) argue that preference instability across methods relates to non-expected utility preferences (Starmer, 2000) and context-dependent preferences (Weber et al., 2002).

Relating to this discussion, Dohmen et al. (2011) find that participants' willingness to take risk varies with context, but are largely correlated. They suggest that elicited risk measures contain a context-specific component, but also a common trait that underlies the responses in the different contexts. In a similar vein, Lévy-Garboua et al. (2012) provide evidence that the rate of inconsistent choices varies for different frames of the same lottery choice experiment (see also Meraner et al., 2018). Deck et al. (2013) do not find evidence that domain specificity explains the observed variation in revealed risk preferences across four elicitation methods and additional survey questions. Relating to the discussion of a stable risk-preference trait, Frey et al. (2017) report experimental evidence that a general factor of risk preferences explains a substantial part of the variance in questionnaires, but less so in experimental methods. Moreover, they report notable inconsistencies in revealed preferences across eight incentivized experimental methods and self-reported questionnaires, with the latter showing more internal consistency (see also Mata et al., 2018).

Alternative explanations of the observed inconsistencies across tasks are provided in a between-subject analysis by Crosetto and Filippin (2015). Even accounting for task-specific measurement errors, they report substantial variation in risk preference estimates across four elicitation methods and discuss potential explanations based on the availability of a safe option and the difference between a single- and a multiple-choice environment. Dulleck et al. (2015) find that between-subject consistency is higher compared to within-subject consistency. Similarly to Crosetto and Filippin (2015), Pedroni et al. (2017) find substantial inconsistency across six risk elicitation mechanisms even when controlling for measurement errors and subjects' numeracy. Furthermore, they do not find support for the assumption that different subjects consistently decide according to Expected Utility or Prospect Theory across tasks. In

a recent experimental study with six elicitation methods, Friedman et al. (2018) find that an expected utility framework decently explains subject behavior in revealing risk preferences except for across-methods inconsistency. The authors further report that some of the inconsistencies can be explained by characteristics of the elicitation methods, such as spatial representation or whether prices or probabilities are varied. Similalry, using two risk elicitation methods by Wakker and Deneffe (1996) and Tanaka et al. (2010), Bauermeister et al. (2017) report inconsistencies in revealed preferences as well as in revealed probability weightings of the lotteries used.

Overall, the reported correlations between risk-taking behavior in different methods tend to be positive and significant, indicating that a certain degree of preference stability cannot be readily dismissed as spurious associations. Moreover, there is some evidence of higher correlations between similar methods (see, e.g., Harrison and Ruström, 2008) and outcomes (see, e.g., Ruggeri and Coretti, 2015). Though, as argued above, it is not clear how to interpret the observed behavior in terms of (in)consistency. It is important to understand whether the literature actually reports on inconsistent behavior or rather on varying preferences of experimental subjects. The primary goal of our study is not to add to the pile of evidence of seemingly inconsistent behavior, but rather to contribute to the understanding of the observed across-method variation in risk preferences. Our main contribution to the literature is to argue that participants in our experiment are well aware of the riskiness associated with their choices and, thus, that their behavior should not be readily interpreted as inconsistent.

3. Experimental Design

We conducted ten experimental sessions with a total of 198 participants (55% female; age: $m = 22.9$ years, $sd = 2.5$) in the Innsbruck EconLab. The experiment was computerized using *oTree* (Chen et al., 2016), utilizing the ready-made applications for risk preference elicitation methods by Holzmeister and Pfurtscheller (2016) and Holzmeister (2017). Participants – bachelor and master students from various fields of study – were recruited using *hroot* (Bock et al., 2014). Upon arrival in the laboratory, participants were seated randomly and asked to start the experiment after carefully reading the instructions on screen. Experimental sessions were conducted in German, took approximately 40 minutes, and were all administered by the same experimenters. Participants received an average payment of €21.35 including a show-up fee of €4.00 (*sd* = €6.25, *min* = €8.00, *max* = €38.50).

We used a within-subject design to measure individual-level risk preferences in four different risk elicitation methods, all of which are commonly applied in economic and social science experiments: (i) the "bomb" risk elicitation task, the certainty equivalent method, a multiple choice list between pairs of lotteries, and a single choice list. Since numerous methods have been introduced to measure risk preferences in the lab, our selection necessarily involves a moment of arbitrariness. However, the four risk preference elicitation tasks included in our study continue to be among the most popular and most widely used ones, despite the fact that correlations between the different measures reported in the literature are usually rather low (Deck et al., 2013). Thus, we deem our choice a good starting point for our analysis.

The parametrization of each task has been mapped to the lottery payoffs and probabilities proposed in the original articles but were scaled in such a way that the expected payoffs of a risk neutral decision maker are similar across tasks ($\sim$ €12.00). The instructions for each of the elicitation methods were

displayed just before participants were asked to make their choice(s) in the particular decision problem. Translated instructions and screenshots of the entire experiment are provided in the Electronic Supplementary Material.

To avoid order and learning effects across tasks (see, e.g., Carlsson et al., 2012), each participant faced a random sequence of the four risk preference elicitation methods.[2] To avoid portfolio-building and cross-task contamination effects (see, e.g., Cubitt et al., 1998; Harrison and Rustrom, 2008), a random lottery incentive system was implemented, i.e., only one of the four tasks was randomly chosen for a subject's final payment (Azrieli et al., 2018).[3] A persistent phenomenon in choice list elicitation procedures is the observation of multiple switching behavior (see, e.g., Bruner, 2011), violating monotonicity and transitivity of revealed preferences and, thus, the paradigm of utility maximization. As our intent is to examine (in)consistency *between* rather than within tasks, we enforced a single switching point in the two multiple price list tasks (CEM and MPL) as proposed by Andersen et al. (2006) and utilized by Jacobson and Petrie (2009) and Tanaka et al. (2010) among others.[4]

3.1. Elicitation methods

In the following, $(x,\, p;\, y)$ denotes a two-outcome lottery that assigns probability p to outcome x and probability $1-p$ to outcome y. Subscripts h and l refer to "high" and "low" lottery outcomes, respectively.

The "bomb" risk elicitation task (BRET). The BRET is a visual risk preference elicitation method requiring subjects to decide on how many boxes to collect out of a matrix containing n boxes. Each box collected yields a payoff γ; but in one of the boxes a "bomb" is hidden, destroying all prospective earnings. Thus, potential earnings increase linearly, but are zero if the bomb is contained in one of the collected boxes. By this means, the BRET elicits (within-method) consistent decisions in $n+1$ lotteries $(\gamma k,\, {}^{(n-k)}/_{n};\, 0)$ and measures individual-level risk attitudes by a single parameter $k \in \{0, 1, \ldots, n\}$, the number of boxes collected. As in Crosetto and Filippin (2013), boxes were collected dynamically and randomly with a time interval of one second for each box once the "Start" button was hit until the "Stop" button was hit.[5] The location of the bomb has only been revealed at the end of the task. In our

[2] Note that, despite a random sequence of tasks, the order in which subjects face the elicitation methods might affect their choices. Thus, we provide a comprehensive analysis of potential order effects in Section A.4 in the Appendix. The results are not indicative of any systematic effects, suggesting that the randomization of tasks on the subject level was an effective means to mitigate potential order effects.

[3] Examining the stability of risk preferences across different methods *on the individual level* calls for a within-subjects experimental design. A within-subject design may induce cross-task contamination effects and necessitates the random lottery incentive system, which effectively introduces a compound lottery. Cubitt et al. (1998) and Starmer and Sugden (1991) provide empirical evidence for the validity of the random lottery incentive system and do not find an indication of contamination effects (see also Harrison and Rustrom, 2008). In line with these results, our analysis of potential order effects (see Section A.4 in the Appendix) does not point towards contaminating effects between tasks in our data.

[4] Note that by enforcing a single switching point, we impose that subjects comply with monotonicity and transitivity requirements, foregoing any opportunity to check whether this is actually the case. Apart from enforcing a single switching point, several alternatives how to deal with multiple switching behavior have been proposed in the literature, such as dropping observations (e.g., Deck et al., 2013), treating the number of safe choices as an indicator of risk preferences (e.g., Holt and Laury, 2002), or adding an indifference option to the choice list (e.g., Andersen et al., 2006).

[5] In Crosetto and Filippin (2013)'s baseline condition "Dynamic," boxes are not collected randomly but sequentially. Our implementation corresponds to their robustness treatment "Random." While the mean number of boxes collected in the "Random" condition is slightly smaller than in the baseline treatment "Dynamic," the distribution of choices across the two treatments does not differ significantly.

experiment, we set n to 100 and γ to €0.50, implying an expected payoff of €12.50 for a risk neutral decision maker.

Certainty equivalent method (CEM). The CEM elicits the point of indifference between a fixed risky lottery L^A = (a_h, p; a_l) with $a_h > a_l$ and n varying degenerated lotteries, i.e., sure payoffs L_i^B = (b_i, 1), with $a_h \leq b_i \leq a_l$ for all $i = 1, 2, \ldots, n$. We implement the parametrization used by Abdellaoui et al. (2011) with $n = 9$ binary choices scaled by a factor of 0.5, i.e., $a_h =$ €15.00, $a_l =$ €5.00, and $b_i = \{$€5.00, €6.25, …, €15.00$\}$. A risk neutral subject expects to earn €11.39.

Multiple price list (MPL). The MPL is characterized by a set of ten binary choices between lotteries with fixed payoffs but varying probabilities of high and low outcomes for each choice. That is, subjects face a menu of n binary choices between lottery L_i^A = (a_h, p_i; a_l) and lottery L_i^B = (b_h, p_i; b_l) for $i = 1, 2, \ldots, n$, where $b_h > a_h > a_l > b_l$. We use the parametrization with $n = 10$ lotteries as proposed by Holt and Laury (2002) but scaled the payoffs by a factor of 5, i.e., $a_h =$ €19.25, $a_l =$ €0.50, $b_h =$ €10.00, and $b_l =$ €8.00 with $p_i = \{0.10, 0.20, \ldots, 1.00\}$. A risk neutral individual expects a payoff of €12.14.

Single choice list (SCL). The SCL offers subjects a menu of different lotteries, asking them to choose the one they prefer to be played. The menu consists of six lotteries which are similar to the implementation proposed by Eckel and Grossman (2002, 2008): L_1 = (€9.00, 0.50; €9.00), L_2 = (€7.50, 0.50; €12.00), L_3 = (€6.00, 0.50; €15.00), L_4 = (€4.50, 0.50; €18.00), L_5 = (€3.00, 0.50; €21.00), and L_6 = (€0.00, 0.50; €24.00). Note that lotteries L_5 and L_6 have the same expected payoff but differ in standard deviation. That is, choosing L_5 implies that the decision maker is either (weakly) risk averse or risk-neutral; choosing L_6 reveals risk neutrality or risk seeking preferences. Hence, a risk neutral decision maker opts either for lottery L_5 or lottery L_6, implying an expected payoff of €12.00.

3.2. Questionnaires

To relate the observed behavior in the four elicitation methods to subjects' perception of the tasks' characteristics as well as their comprehension and numeracy, the experimental protocol comprised several additional questionnaires. Details on the questionnaires and subjects' responses are provided in Sections A.1, A.2, and A.3 in the Appendix. Note that the survey items were not incentivized. Thus, our approach of combining experimental with questionnaire data is somewhat exploratory in nature. However, given the vast number of puzzling findings on the instability of risk preferences in the literature and the lack of a consistent interpretation thereof, such an exploratory approach can be useful to shed some light on potential mechanisms driving across-methods instability.

Directly after a decision in any of the four tasks has been submitted, participants were asked to assess how risky they perceive their decision to be and how confident they feel about the particular choice they made. Each decision was depicted, as participants have just completed it, on a separate screen and questions were answered on a scale from 1 ("not at all risky/confident") to 7 ("very risky/confident").[6] On the premise that subjects' risk preferences are a stable trait, one would expect to observe identical—or at least similar—assessments of the riskiness of choices across the four tasks on the individual level.

[6] Note that enforcing a single switching point in the two multiple choice list tasks (CEM and MPL) as proposed by Andersen et al. (2006) might affect how participants qualitatively evaluate the risk taken as well as their confidence. Yet, we cannot think of a particular argument for systematic effects of enforced within-task consistency on subjects' self-assessed risk-taking and confidence.

To examine whether insufficient comprehension of the elicitation procedures gives rise to increased across-methods variation in revealed risk preferences, the experimental protocol included comprehension questions and an 8-item Rasch-validated numeracy inventory (Weller et al., 2013). For the comprehension questions, subjects were shown a screenshot of the risk neutral decision in each of the four tasks and were asked to estimate (i) the expected payoff, (ii) the probability to earn less than €5.50, and (iii) the probability to earn more than €14.50. Given the assumption that participants' choices are dictated by some latent, deterministic preference relation, mistakes in evaluating the available lottery choices might impair across-methods consistency. We, thus, conjecture that the likelihood of making mistakes is negatively related to subject's numeracy and comprehension of tasks. Accordingly, we expect to observe a negative relation between across-methods preference variation and comprehension and numeracy, respectively.

Moreover, we elicited several qualitative judgments on how subjects perceive the tasks relative to the other methods. After completing all elicitation methods, subjects were therefore presented with additional questionnaires, requiring them to explicitly compare the four elicitation methods with regards to various dimensions on a single screen. In particular, we asked participants to evaluate each of the four elicitation methods with respect to (i) whether the instructions are easy to understand, (ii) whether answering the task involves complex calculations, (iii) whether the task is boring, and (iv) whether the decision problem is associated with an investment, gambling, or insurance domain. Each of the questions (i) to (iii) was answered on a scale from 1 ("not agree at all') to 7 ("fully agree"). For answering question (iv), subjects had to indicate whether they associate the task with the investment, gambling, or insurance domain using a drop-down field. We hypothesize to find more noisy behavior within tasks that are perceived to be complex. Furthermore, subjects' association with a specific domain serves as a means to examine whether revealed risk preferences are domain-specific. We conjecture to find less variation in revealed preferences for elicitation methods that are assigned to the same domains compared to elicitation methods that are associated with different domains.

4. Analysis framework

For the analysis of the experimental data, we assume an expected utility theory (EUT) framework. To estimate risk preferences, we assume a standard isoelastic utility function—a member of the family of power utility functions—of the form

$$u(x) = \begin{cases} (1-\varphi)^{-1}\, x^{1-\varphi} & \text{if } \varphi \neq 1 \\ ln(x) & \text{if } \varphi = 1 \end{cases} \tag{1}$$

which is characterized by constant relative risk aversion (CRRA). This specification of utility curvature has been widely used in economics and related fields and has been shown to typically better fit experimental data than alternative families (Camerer and Ho, 1994; Wakker, 2008).

To examine whether variation in revealed preferences is correlated with explanatory measures elicited in the questionnaires, an individual-level measure of the across-methods stability of revealed preferences is required. Note that the assumption of a parametric functional form of a participant's utility function implies that observed choices in a risk preference elicitation method translate into parameter

intervals rather than point estimates. We define choices in two independent tasks as "stable" if the implied parameter intervals overlap (see, e.g., Bruner, 2009). Whenever the sets of feasible parameters implied by the choices in two methods intersect, it cannot be ruled out that the observed choices do indeed stem from the same latent parameter φ. In particular, we define an indicator for each pairwise comparison of methods, which is equal to one if the implied parameter intervals overlap, and zero otherwise. As a preference stability index, we sum up these binary indicators for all six unique pairwise combinations of the four experimental risk preference elicitation methods, implying a measure between 0 and 6 on the individual level. This measure is conservative for two reasons: First, overlapping parameter intervals do not necessarily imply identical risk aversion parameters and, thus, across-methods invariance in the sense of risk preferences as a stable trait. Second, an overlap of parameter intervals could eventually be the result of random behavior or chance. For these reasons, the index has to be interpreted as a proxy for preference invariance.

In addition to the individual-level preference stability index we examine across-methods variation of risk preferences on the aggregate level by estimating a structural model for each elicitation method. We follow the procedure for structural model estimation for binary discrete choices under risk discussed in Harrison and Rustrom (2008) and Wilcox (2008). Given the assumption of an EUT framework, the probabilities p_k for the high and low lottery payoffs $k \in \{h, l\}$ are those that are induced in the particular elicitation method by the experimenter. Thus, the expected utility of lottery i, $E[u_i]$, is the utility of each lottery outcome, u_k, weighted by the corresponding probability:

$$E[u_i] = \sum_k p_k u_k \quad \forall\, k \in \{h, l\} \tag{2}$$

For each of the $i = 1, 2, \ldots, n$ lottery pairs, participants are assumed to choose either the less risky (or safe) lottery A or the more risky lottery B by evaluating the difference between their expected utilities.[7] In addition, we allow for mistakes or trembles in comparing the expected utilities of the alternatives participants face, modeled as a *Fechner* error term (see, e.g., Hey and Orme, 1994; Loomes et al., 2002), yielding the latent index

$$\nabla E[u_i] = E[u_B] - E[u_A] + \sigma\epsilon \quad \text{with } \epsilon \sim N(0, 1) \tag{3}$$

The additive component $\sigma\epsilon$ is a stochastic error term and can be interpreted as capturing noise in the decision maker's evaluation of the difference between the lotteries' expected utilities, with σ being proportional to the standard deviation of this noise (Wilcox, 2008).

The index $\nabla E[u_i]$, determined by latent preferences, is then linked to the participants' observed choices using the cumulative standard normal distribution $\Phi(\cdot)$.[8] This implies that the latent variable model of

[7] In order to apply this procedure, choices in all elicitation methods need to be expressed as a series of binary choices between lottery pairs. While this is the case for the CEM and the MPL by default, data from the BRET and the SCL need to be transformed. Following Dave et al. (2010) and Crosetto and Filippin (2015), we convert the gambles in BRET and SCL into implicit binary choices between two adjacent gambles assuming that utility functions are well-behaved, i.e., that preferences are single-peaked. Thus, for the BRET, for instance, a subject selecting 40 out of 100 boxes is assumed not only to reveal that 40 boxes are preferred to 39 but also that 39 boxes are preferred to 38, 40 boxes are preferred to 41, etc. The same rationale is applied to the observed choices in the SCL.

[8] Alternatively, the probit link could be replaced by a logit link as proposed by Luce and Suppes (1965) and employed by Camerer and Ho (1994) and Dave et al. (2010), among others. For our data, the results turn out to be qualitatively akin for either of the two functional specifications.

a considered choice probability using a probit link function is given by

$$\begin{aligned} P(B \succ A) &= \Phi\left(\nabla E[u_i]\right) \\ P(B \succ A) &= \Phi\left(\sigma^{-1}\left(E[u_B] - E[u_A]\right)\right) \end{aligned} \tag{4}$$

That is, the latent index $\nabla E[u_i]$ is linked to the observed choices by the specification that lottery B is chosen whenever $\Phi(\nabla E[u_i]) > 1/2$. As the standard deviation of the structural noise term, σ, approaches zero, the probability that the observed choice reflects the latent preference relation converges towards one.

The likelihood of participants' responses, $L(\cdot)$, thus, is a function of the CRRA parameter φ, the standard deviation of the structural noise σ, and the vector of n choices observed in the experimental task ($\vec{y}$). The conditional log-likelihood function is given by

$$\ln L(\varphi, \sigma|\vec{y}) = \sum_{i=1}^{n} \left(\left[\ln \Phi(\nabla E[u_i]) \right]^{y_i} + \left[\ln \Phi(-\nabla E[u_i]) \right]^{1-y_i} \right) \tag{5}$$

where y_i denotes an indicator function taking value 1 if a participant chooses the more risky lottery B and zero otherwise, for all $i = 1, 2, \ldots, n$. The function $\ln L(\varphi, \sigma|\vec{y})$ is maximized with respect to φ and σ, with standard errors being clustered on the subject level, reproducing the routines for *Stata* proposed by Harrison and Ruström (2008).

At this point it should be noted that random utility models, which include the model delineated above, have recently been shown to be prone to violations of monotonicity. In particular, the choice probability $P(B \succ A)$ is not necessarily a decreasing function of the CRRA parameter φ, whereas random parameter models are always monotone in this regard (Apesteguia and Ballester, 2018).[9] However, in our specific setting, the methodology of the random parameter model has disadvantages, which are discussed in detail in Section A.5 in the Appendix. For this reason, we assume a random utility model in our analysis and only refer to the alternative model specification where relevant.

5. Results

In what follows we first present evidence on the instability of risk preferences, then relate it to subjects' perceived riskiness of choices, and finally discuss potential explanations of our findings in the light of the related literature.

5.1. Instability of Risk Preferences

In line with previous results on across-methods variation in risk preferences (see, e.g., Csermely and Rabas, 2016; Deck et al., 2013; Dulleck et al., 2015; Pedroni et al., 2017), we find that Spearman rank correlations between the observed number of risky choices in the four tasks are moderate but significantly different from zero, varying between 0.157 and 0.326; polychoric correlations are slightly higher

[9] In particular, the use of random utility models may pose identification problems and could yield biased estimates, since the same probabilities of choosing the risky alternative B may be associated with different levels of risk aversion. For a detailed discussion see Apesteguia and Ballester (2018).

and vary between 0.189 and 0.396 (Tab. 1). While compared to other results reported in the literature (see e.g., Deck et al., 2013; Pedroni et al., 2017) we find moderately higher and consistently significant correlations, overall they are still rather low in magnitude. This indicates that the ranking across subjects is not preserved across different methods. Indeed, only 71.7% of the participants are consistently risk averse in all four tasks. For the remaining 28.3% of the participants, choices are associated with risk loving preferences at least once.

Table 1: Correlation matrix. The lower triangular matrix reports Spearman rank correlations between the observed number of risky choices in the four tasks; the upper triangular matrix depicts polychoric correlations. p-values are reported in parentheses ($n = 198$). BRET, CEM, MPL, and SCL denote the "bomb" risk elicitation task, the certainty equivalent method, the multiple price list, and the single choice list, respectively.

	BRET	CEM	MPL	SCL
BRET		0.245	0.350	0.336
		(0.001)	(0.000)	(0.000)
CEM	0.222		0.283	0.400
	(0.002)		(0.000)	(0.000)
MPL	0.367	0.244		0.387
	(0.000)	(0.001)		(0.000)
SCL	0.341	0.338	0.354	
	(0.000)	(0.000)	(0.000)	

With respect to across-methods stability of preferences, subjects on average reveal stable risk preferences in 2.8 ($sd = 1.5$) out of 6 possible combinations.[10] In order to appropriately interpret the degree of observed variation in preferences, it is informative to relate the experimental data to sensible benchmarks. The theoretical upper bound of the preference stability index is derived from a hypothetical subject with deterministic and stable preferences who does not make any mistakes in revealing her preferences in any of the tasks. Such a subject would act exactly as her φ dictates and reveal invariant preferences in all six pairwise comparisons in our setting. As the sets of feasible CRRA interval estimates implied by participants' choices in the elicitation methods might intersect by pure chance, even random behavior can be expected to manifest itself in a preference stability index larger than zero. To approximate a sensible lower benchmark we thus simulate uniformly distributed choices for each of the four methods for 10,000 virtual subjects characterized by the preference functional as described above. Indeed, these simulations reveal that the lower benchmark is substantially larger than zero ($m = 1.3$, $sd = 1.1$), with only $\sim 1/4$ of the simulation outcomes ending up with 0 out of 6 possible intersections of CRRA point estimate sets. Two more simulation exercises are informative as sensible benchmarks for the experimental data. In the first simulation, choices for each of the four tasks are drawn *independently* from the choice distribution observed in the laboratory data. By that means, the simulation exercise assumes that subjects treat each of the tasks independently. An alternative bench-

[10] BRET, MPL, and CEM include at least one first-order dominated choice each. Of the 198 subjects in our sample, 13 (6.6%) violate basic rationality by choosing a dominated lottery in at least one of the tasks: 1 (0.5%) in BRET, 6 (3.0%) in CEM, and 9 (4.5%) in MPL. As dominated choices cannot be translated into CRRA intervals, the preference stability index cannot be reasonably determined for participants violating rationality. Thus, any result referring to the preference stability index is based on the sample with $n = 185$.

mark, motivated by Crosetto and Filippin (2015), is determined by virtual subjects exhibiting stochastic preferences. For this purpose, we simulate another 10,000 virtual subjects characterized by some latent CRRA parameter φ_l but add some i.i.d. noise directly to subject's inherent risk preferences for each of the four methods. In particular, we assume that the virtual subjects' latent parameter φ_l is normally distributed, with $\mu_l = 0.6$ and $\sigma_l = 0.3$.[11] That is, the actual φ_a determining virtual subject's choices departs from their real, latent φ_l by some stochastic noise with zero mean and standard deviation σ_a, i.e. $\varphi_a = \varphi_l + \sigma_a, \sigma_a \sim N(0, 0.2)$.

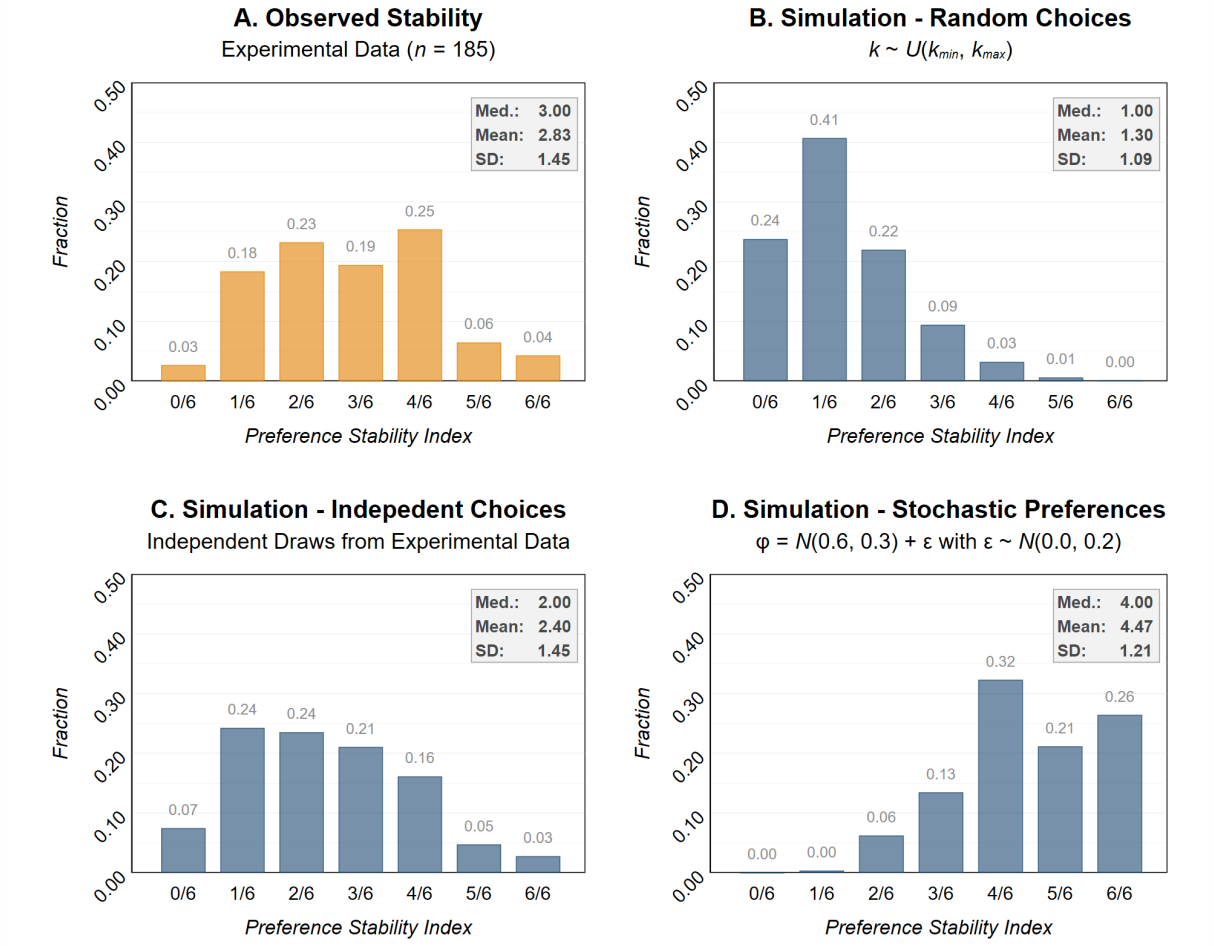

Figure 1: (A) Distribution of preference stability (number of pairwise comparisons in which implied parameter intervals overlap) for the experimental data ($n = 185$). **(B)** Simulation exercise with virtual subjects choosing uniformly and independently from the available choices in each of the four risk preference elicitation methods. **(C)** Simulation exercise with virtual subjects choosing independently from the choice distribution of each task as observed in the experiment. **(D)** Simulation exercise with virtual subjects with stochastic preferences, where a noise term $\epsilon \sim N(0, 0.2)$ is added directly to subjects' CRRA parameter $\varphi \sim N(0.6, 0.3)$. $n = 10,000$ for each simulation.

The distributions of preference stability indices observed in the experiment as well as the three simulation results are depicted in Figure 1. Eyeballing the histograms indicates that the distribution from the experimental data (panel A) can neither be fully explained by subjects choosing uniformly at random

[11] The values for μ_l and σ_l are similar to aggregate maximum likelihood estimates for our sample such that the distribution of the deterministic part in the simulation resembles the overall mean and variance in risk preferences in the experiment.

(panel B), nor by subjects characterized by stochastic preferences (panel D). The simulation outcomes of independent draws from the experimental data (panel C), however, highlight considerable similarities to the experimental data. This is a surprising result, as the observed distribution in the laboratory reveals a behavioral pattern that appears as if experimental subjects would choose *independently* across the four elicitation methods. Despite the observed significant correlations between risky choices across tasks, these results raise the question *why* participants exhibit such a high level of variation in revealing their risk preferences.[12]

5.2. Perceived Riskiness of Choices

One intuitive interpretation of the above finding is that any observed heterogeneity in revealed risk attitudes results from *inconsistent* behavior. This implicitly assumes that subjects would want to choose as dictated by a stable risk preference relation, but are either unaware of the actual variation in their risk-taking behavior or simply unable to make choices that reflect these stable preferences across tasks. However, in what follows, we provide evidence that subjects deliberately make choices characterized by varying risk preferences across tasks.

On the aggregate level, we estimate structural models for each of the tasks, as described in Section 4. The corresponding maximum likelihood estimates, $\hat{\varphi}$ and $\hat{\sigma}$, are reported in Tab. 2A. Estimates of both the CRRA coefficient and the variance of noise differ considerably across methods. The CRRA estimates for all pairwise comparisons of methods are significantly different from one another, except for $\hat{\varphi}_{\text{BRET}}$ and $\hat{\varphi}_{\text{MPL}}$ (lower triangular matrix in Tab. 2B). Differences between the estimates of the variance of the structural noise term are statistically significant for all comparisons of methods and show even more pronounced effect sizes (upper triangular matrix in Tab. 2B). Note that the maximum likelihood estimates of the CRRA parameter φ are by all means comparable to estimates reported in the literature in terms of magnitude. In particular, we are not the first to report that subjects, on average, tend to be significantly more risk averse in the BRET and the MPL than in the SCL(see, e.g., Crosetto and Filippin, 2015; Dave et al., 2010).

Comparing CRRA point estimates $\hat{\varphi}$ (Fig. 2A) to the average subject-level demeaned perceived riskiness of each task (Fig. 2B) reveals a remarkable result. Not only do the assessments of riskiness differ considerably across tasks, but the almost perfectly mirrored patterns suggest that, on average, subjects are well aware of the level of and the across-methods variation in the riskiness associated with their choices. This is a strong indicator that subjects *deliberately* take different levels of risk across tasks. This awareness even extends to the participants' assessment of the difficulty of tasks. Panels C and D of Fig. 2 depict maximum likelihood estimates of the standard deviation of the noise parameter σ in the structural model for each elicitation method as well as the average subject-level demeaned perception of the tasks' complexity. Again, both patterns look similar to a remarkable extent, indicating that

[12] Please note that there might be an effect of the different sizes of implied intervals of CRRA parameters φ on the preference stability index. Both the structure and the size of the implied intervals is determined by the structure of payoffs and probabilities of the respective lotteries, and as such differ considerably across the four elicitation procedures examined. Assuming that the choice architecture of a particular task may systematically affect subjects' risk-taking behavior, the choice itself may also systematically affect the implied interval size and, in turn, subjects' stability indices. Hence, the relationship between the interval size and participants' susceptibility to making mistakes in revealing their actual risk preferences is not straightforward or clearly unidirectional.

Table 2: (A) Maximum likelihood estimates of structural models with *Fechner* error terms for each of the four risk preference elicitation methods. Standard errors, clustered on the subject level, are reported in parentheses. **(B)** Pairwise differences in point estimates of risk preference parameters φ (lower-triangular matrix) and the standard deviation of noise parameters σ (upper-triangular matrix) between the four risk preference elicitation methods. p-values are based on pairwise Wald tests. BRET, CEM, MPL, and SCL denote the "bomb" risk elicitation task, the certainty equivalent method, the multiple price list, and the single choice list, respectively. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Panel A	BRET	CEM	MPL	SCL
φ	0.626^{***}	0.838^{***}	0.602^{***}	0.387^{***}
	(0.021)	(0.090)	(0.033)	(0.034)
σ	0.046^{***}	0.263^{***}	0.977^{***}	0.720^{***}
	(0.002)	(0.048)	(0.066)	(0.057)
$\ln L$	−5,298	−458	−600	−572
No. of Obs.	19,800	1,782	1,980	990
Clusters	198	198	198	198

Panel B	BRET	CEM	MPL	SCL
BRET		-0.217^{***}	-0.932^{***}	-0.674^{***}
CEM	0.212^{*}		-0.715^{***}	-0.457^{***}
MPL	−0.025	-0.237^{**}		0.257^{**}
SCL	-0.240^{***}	-0.452^{***}	-0.215^{***}	

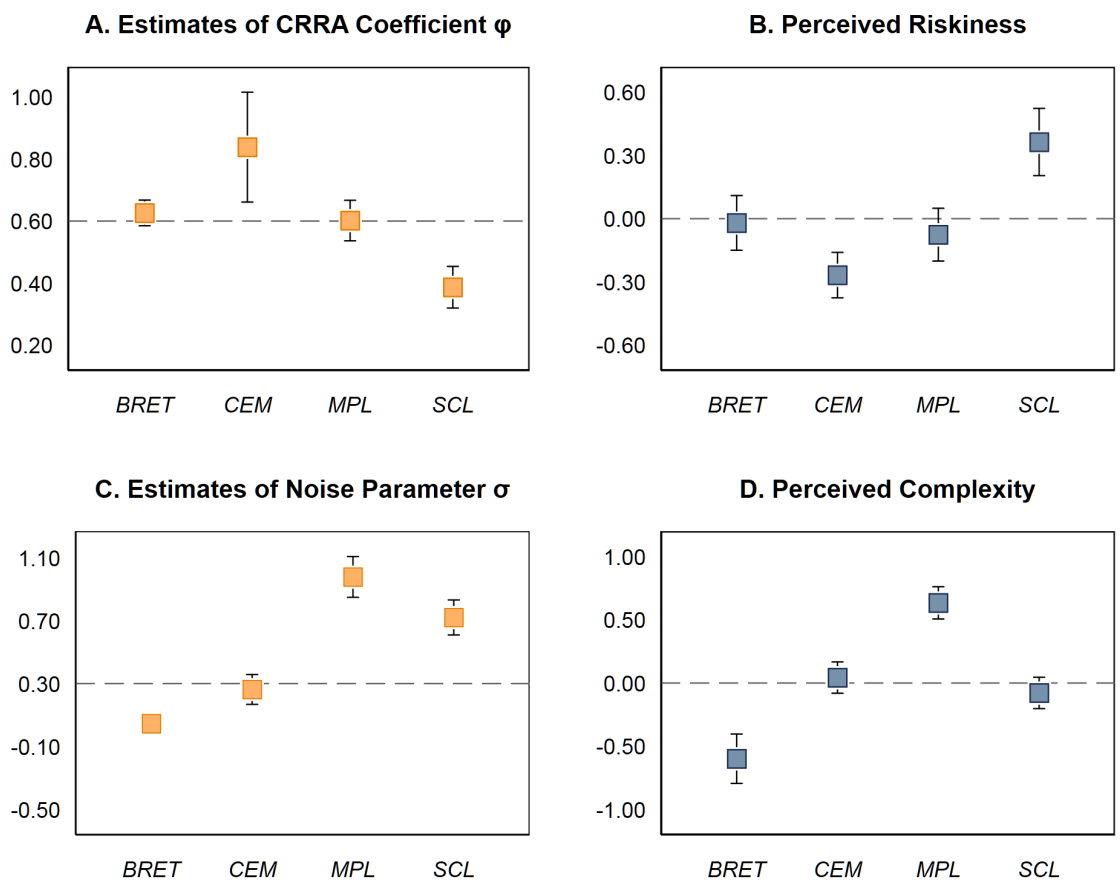

Figure 2: (A) Maximum likelihood estimates of CRRA coefficients φ. **(B)** Average perceived riskiness (subject-demeaned data) for the four risk preference elicitation methods. **(C)** Maximum likelihood estimates of the standard deviation of the structural noise parameter σ. **(D)** Average perceived complexity (subject-demeaned data) for the four risk preference elicitation methods. In all panels, error bars indicate 95% confidence intervals; dashed lines approximate the overall estimate in panels A and C ($\hat{\varphi} = 0.585$ and $\hat{\sigma} = 0.324$) and depict means in panels B and D. Standard errors in the maximum likelihood estimations are clustered on the individual level; $n = 198$. BRET, CEM, MPL, and SCL denote the "bomb" risk elicitation task, the certainty equivalent method, the multiple price list, and the single choice list, respectively.

subjects, on average, can well assess the susceptibility to mistakes or trembles in revealing their actual preferences across methods.[13]

We provide additional evidence on subjects' awareness of varying levels of risk associated with seemingly inconsistent choices across methods by extending the structural model specification outlined in section 4. In particular, we estimate $\hat{\varphi} = \hat{\varphi}_0 + \hat{\varphi}_r \cdot r_p$ and $\hat{\sigma} = \hat{\sigma}_0 + \hat{\sigma}_c \cdot c_p$, where $\hat{\varphi}_0$ and $\hat{\sigma}_0$ are estimates of the constants and r_p and c_p refer to perceived (subject-level demeaned) riskiness and complexity, respectively. The maximum likelihood estimates of risk aversion in this model significantly correlates with participants' evaluation of the choice's riskiness ($\hat{\varphi}_r = -0.131$, $p < 0.001$), and the variance of the structural noise term significantly varies depending on subjects' appraisal of task complexity ($\hat{\sigma}_c = 0.065$, $p < 0.001$). Overall, our results indicate that subjects seem to be well aware of the

[13] It is reassuring that the estimates of φ based on a random parameter model, reported in Table S5 in the Appendix, are qualitatively similar to the results of the random utility model reported in Table 2. In particular, the ordering of point estimates is preserved and that patterns of significant differences remain similar using the alternative model specification.

riskiness of their choices as well as the complexity of the decision situation.

Our findings are in line with the observed zero correlation of (i) numeracy and (ii) task comprehension with preference stability in our experimental data: We hypothesized that subjects' ability to reveal their risk preferences may vary across the different elicitation methods. Subjects might make mistakes in evaluating the lotteries that are explicitly and implicitly contained in the elicitation procedures, and thus in correctly choosing the lotteries that match their preferences. Accordingly, we should find a correlation between a subjects' level of preference stability and (i) the absolute difference between subjects' responses and the correct answers to the comprehension questions[14], and (ii) the achieved numeracy scale. However, both correlations are low and insignificant ($\varrho = -0.089, p = 0.210$ and $\varrho = 0.033, p = 0.649$, respectively). Thus, we do not find evidence of a positive relation between a subject's numeracy or comprehension of tasks and the degree of preference stability across tasks.[15] We deem this finding anything but trivial. It corroborates the basic assumption that risk preference elicitation methods are indeed designed in a way that subjects are able to reveal their preferences irrespective of their explicit understanding of the calculations behind the lotteries. Moreover, this zero correlation is in line with our conclusion that subjects are well aware of the difficulty of methods and the susceptibility to mistakes, but still make choices that differ in the riskiness across tasks.

At this point it could be argued that we cannot rule out that subjects are inconsistent. They might be unaware of being driven to varying decisions by the design of a particular method. Such unawareness, however, is not in line with our data, since ignorant subjects with stable risk preferences would have to assess their decisions as equally risky in each method. In contrast, subjects might not be able to evaluate whether they are *systematically* driven to more or less risky decisions by the different elicitation procedures. However, in this case subjects' risk assessments should be noisy, which, again, contradicts the systematic differences in reported riskiness across methods that we find. Moreover, we see no reason to assume that subjects make their choices according to the methods' systematic drive, which they are aware of and which contradicts their own preferences.

Therefore, we conclude that subjects *deliberately* make choices that reveal varying risk attitudes across methods.[16] While this result by all means challenges the procedural invariance axiom, there are several possible interpretations.

5.3. Possible Explanations

One potential explanation of the variation in risk attitudes across methods reported above is that subjects do not behave upon the same risk preference relation in different elicitation methods. In particular, subjects might have domain-specific risk preferences for different types of choices (Weber et al., 2002). To account for this possibility, we elicited subjects' association of methods with an investment, gambling, or insurance domain. For pairwise comparisons of methods, we test if the preference stability

[14] For each of the three questions per task, we first calculate the absolute difference between subjects' responses and the correct answers. In a second step we relate each deviation to the correct answer and average them on the subject level. For a comparison of relative absolute deviations per task see Section A.1 in the Appendix.

[15] This is in line with previous literature, such as Pedroni et al. (2017). See also Andersson et al. (2016) and Andersson et al. (2018), who find that cognitive ability is related to noisy behavior rather than risk preferences.

[16] Note that this conclusion is in line with the finding in Dulleck et al. (2015), where only 8 out of 78 subjects wanted to change their decision when given the chance to do so.

index is higher for subjects that assign the same domain to the two tasks compared. As reported in Table S2 in the Appendix, we do not find a significant effect for any of the pairwise comparisons. Although we have a rather crude measure of domain-specificity, with only three choice-options for associated domains, our result is in line with previous findings (see, e.g., Deck et al., 2013). Thus, we cannot conclude that domain-specific preferences are the main driver of the observed variation in revealed risk preferences. Given that our choice of domains is motivated by real-world contexts, i.e., investment, gambling, and insurance, our finding also relates to recent evidence that calls into question the external validity of experimental measures of risk preferences (see Charness et al., 2019).[17]

Alternatively, each of the four methods might elicit a set of different preferences, which subjects are balancing in their choices. For instance, in the BRET the worst outcome, i.e. the minimum gain, for a subject is to earn €0. This is in strong contrast to the other tasks, especially the CEM and the SCL where, in the worst case, subjects cannot fall below €5 and €9, respectively (Crosetto and Filippin, 2015). A subject might make less risky decisions in a choice environment that is perceived as more risky (He and Hong, 2017), because her risk assessment is influenced by the possible worst outcome in the task (Anzoni and Zeisberger, 2016; Holzmeister et al., 2018). In a similar manner, the choice structure of tasks might influence risk-taking behavior, e.g., because subjects tend to avoid extreme choices in the opportunity sets. Examples are provided by Andersen et al. (2006) showing that the available lotteries affect choices, and by Crosetto and Filippin (2017) showing that removing choices in tasks influences risk-taking.

Another conceivable explanation refers to the general framework of analysis. Possibly, expected utility is not the most appropriate framework to interpret subjects' preferences. Rather, they might have reference point-dependent preferences, comprising loss or disappointment aversion (see, e.g. Gul, 1991; Kahneman and Tversky, 1979), that reveal themselves in the observed choices.[18]

Eventually, given our data, we have to remain agnostic about the plausibility of some of the explanations discussed above and further examination has to be left for future research.

6. Conclusion

We conducted an within-subjects experiment with 198 subjects, examining revealed risk preferences in four different, widely used risk preference elicitation methods. In line with previous studies, we find substantial variation in revealed risk preferences. On average, subjects' risk preferences are consistent in less than half of the pairwise comparisons of methods. Comparing the observed behavior to results from simulation exercises, we find that the observed heterogeneity in risk preferences across tasks looks similar to the heterogeneity in independent random draws from choices in the experiment, despite significantly positive correlations between risky choices across tasks. As a novel contribution, we relate the observed behavior to subjects' perceived riskiness of choices reported in a questionnaire.

[17] However, for evidence on the external explanatory power of incentivized measures see, e.g., Anderson and Mellor (2008) and Lusk and Coble (2005); for survey based measures see, e.g., Barsky et al. (1997), Beauchamp et al. (2017), and Dohmen et al. (2011).

[18] Carbone and Hey (1995) argue that the preference functional that can explain subjects' choices may be conditional on the elicitation method. However, recent evidence suggests that the elicitation of risk attitudes is more sensible to the method used than the assumed preference functional (Zhou and Hey, 2017). In line with these results, Pedroni et al. (2017) and Friedman et al. (2018) do not find evidence for superior alternative explanatory frameworks. See also Vosgerau and Peer (2018) for evidence of malleability of preferences under uncertainty.

Notably, subjects seem to be well aware of the level of risk associated with their choices, even though the observed behavior can be characterized by varying risk attitudes. We interpret this as a piece of evidence that participants make their choices *deliberately*. This suggests that subjects' behavior cannot be readily interpreted as inconsistent and that the standard assumption of procedural invariance should be reconsidered.

Our results have several implications: First, they shed light on previous findings on within- as well as between-subject variation of revealed risk preferences across different elicitation methods, in that observed behavior might not be easily dismissed as inconsistency. Second, our results call for a re-assessment of the common research practice of choosing among different elicitation procedures based on purely pragmatic reasons. Similar to the results reported by (Loomes and Pogrebna, 2014; Zhou and Hey, 2017), our findings indicate that the choice of the elicitation method may well have a major impact on the elicited preferences. Eventually, we hope that our results contribute to a fruitful discussion on across-methods stability of revealed risk preferences and the methodology of risk preference elicitation in general.

References

Abdellaoui, M., Driouchi, A., & L'Haridon, O. (2011). Risk aversion elicitation: Reconciling tractability and bias minimization. *Theory and Decision*, *71*, 63–80.

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

Anderson, L. R. & Mellor, J. M. (2008). Predicting health behaviors with an experimental measure of risk preference. *Journal of Health Economics*, *27*(5), 1260–1274.

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

Andersson, O., Holm, H. J., Tyran, J.-R., & Wengström, E. (2016). Risk aversion relates to cognitive ability: Preference or noise? *Journal of the European Economic Association*, *14*(5), 1129–1154.

Andersson, O., Holm, H. J., Tyran, J.-R., & Wengström, E. (2018). Robust inference in risk elicitation tasks. *Working Paper*.

Anzoni, L. & Zeisberger, S. (2016). What is risk? How investors perceive risk in return distributions. *Working Paper*.

Apesteguia, J. & Ballester, M. A. (2018). Monotone stochastic choice models: The case of risk and time preferences. *Journal of Political Economy*, *126*(1), 74–106.

Arrow, K. J. (1965). *Aspects of the theory of risk bearing* (Y. J. Saatio, Ed.). Helsinki: Yrjö Jahnssonin Säätiö.

Azrieli, Y., Chambers, C. P., & Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, *126*(4), 1472–1503.

Barsky, R., Juster, F., Kimball, M., & Shapiro, M. (1997). Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study. *Quarterly Journal of Economics*, *112*(2), 537–579.

Bauermeister, G.-F., Hermann, D., & Musshoff, O. (2017). Consistency of determined risk attitudes and probability weightings across different elicitation methods. *Theory and Decision*, *online first*, 1–18.

Beauchamp, J., Cesarini, D., & Johannesson, M. (2017). The psychometric and empirical properties of measures of risk preferences. *Journal of Risk and Uncertainty*, *54*, 203–237.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, *9*(3), 226–232.

Berg, J., Dickhaut, J., & McCabe, K. (2005). Risk aversion elicitation: Reconciling tractability and bias minimization. *Proceedings of the National Academy of Science of the United States of America*, *102*(11), 4209–4214.

Binswanger, H. P. (1980). Attitudes toward risk: Experimental measurement in rural india. *American Journal of Agricultural Economics*, *62*(3), 395–407.

Binswanger, H. P. (1981). Attitudes toward risk: Theoretical implications of an experiment in rural india. *The Economic Journal*, *91*(364), 867–890.

Bock, O., Baetge, I., & Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, *71*, 117–120.

Bruner, D. M. (2009). Changing the probability versus changing the reward. *Experimental Economics*, *12*(4), 367–385.

Bruner, D. M. (2011). Multiple switching behaviour in multiple price lists. *Applied Economics Letters*, *18*(5), 417–420.

Camerer, C. F. & Ho, T.-H. (1994). Violations of the betweenness axiom and nonlinearity in probability. *Journal of Risk and Uncertainty*, *8*(2), 187–196.

Carbone, E. & Hey, J. D. (1995). A comparison of the estimates of expected utility and non-expected-utility preference functionals. *The Geneva Papers on Risk and Insurance Theory*, *20*(1), 111–133.

Carlsson, F., Mørkbak, M. R., & Olsen, S. B. (2012). The first time is the hardest: A test of ordering effects in choice experiments. *Journal of Choice Modelling*, *5*(2), 19–37.

Charness, G., Gneezy, U., & Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization*, *87*, 43–51.

Charness, G., Offerman, T., T. Garcia, & Villeval, M. (2019). Do measures of risk attitudes in the laboratory predict behavior under risk in and outside of the laboratory? *IZA Discussion Paper, No. 12395*.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Cohen, M., Jaffray, J.-Y., & Said, T. (1987). Experimental comparison of individual behavior under risk and under uncertainty for gains and for losses. *Organizational Behavior and Human Decision Processes*, *39*, 1–22.

Crosetto, P. & Filippin, A. (2013). The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, *47*, 31–65.

Crosetto, P. & Filippin, A. (2015). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics*, *18*(6), 1–29.

Crosetto, P. & Filippin, A. (2017). Safe options induce gender differences in risk attitudes. *IZA Discussion Paper*, *No. 10793s*.

Csermely, T. & Rabas, A. (2016). How to reveal people's preferences: Comparing time consistency and predictive power of multiple price list risk elicitation methods. *Journal of Risk and Uncertainty*, *53*(2), 107–136.

Cubitt, R. P., Starmer, C., & Sugden, R. (1998). On the validity of the random lottery incentive system. *Experimental Economics*, *1*, 115–131.

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.

Deck, C., Lee, J., Reyes, J. A., & Rosen, C. C. (2013). A failed attempt to explain within subject variation in risk taking behavior using domain specific risk attitudes. *Journal of Economic Behavior & Organization*, *87*, 1–24.

Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review*, *100*(3), 1238–1260.

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

Dulleck, U., Fooken, J., & Fell, J. (2015). Within-subject intra- and inter-method consistency of two experimental risk attitude elicitation methods. *German Economic Review*, *16*, 104–121.

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

Eckel, C. C. & Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization*, *68*, 1–17.

Eliashberg, J. & Hauser, J. R. (1985). A measurement error approach for modeling consumer risk preference. *Management Science*, *31*(1), 1–25.

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

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, *3*, e1701381.

Friedman, D., Habib, S., James, D., & Crockett, S. (2018). Varieties of risk elicitation. *WZB Discussion Paper, No. SP II 2018-501.*

Gul, F. (1991). A theory of disappointment aversion. *Econometrica*, *59*(3), 667–686.

Harrison, G. W. & Rustrom, E. E. (2008). Risk aversion in the laboratory. In J. Cox & G. Harrison (Eds.), *Risk aversion in experiments* (pp. 41–196). Research in Experimental Economics 12. Bingley, UK: Emerald.

He, T.-S. & Hong, F. (2017). Risk breeds risk aversion. *Experimental Economics*, *21*(4), 815–835.

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

Hey, J. D. & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica*, *62*(6), 1291–1326.

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

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

Holzmeister, F. (2017). oTree: Ready-made apps for risk preference elicitation methods. *Journal of Behavioral and Experimental Finance*, *16*, 33–38.

Holzmeister, F., Huber, J., Kirchler, M., Lindern, F., Weitzel, U., & Zeisberger, S. (2018). What drives risk perception? A global survey with financial professionals and lay people. *OSF Preprints*. doi:10.31219/osf.io/v6r9n

Holzmeister, F. & Pfurtscheller, A. (2016). oTree: The "bomb" risk elicitation task. *Journal of Behavioral and Experimental Finance*, *10*, 105–108.

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

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

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Lévy-Garboua, L., Maafi, H., Masclet, D., & Terracol, A. (2012). Risk aversion and framing effects. *Experimental Economics*, *15*, 128–144.

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

Loomes, G. & Pogrebna, G. (2014). Measuring individual risk attitudes when preferences are imprecise. *The Economic Journal*, *124*(576), 569–593.

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

Luce, R. D. & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. B. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 3, pp. 249–410). New York: Wiley.

Lusk, J. & Coble, K. (2005). Risk Perceptions, Risk Preference, and Acceptance of Risky Food. *American Journal of Agricultural Economics*, *87*(2), 393–405.

Mata, R., Frey, R., Richter, D., Schupp, J., & Hertwig, R. (2018). Risk reference: A view from psychology. *Journal of Economic Perspectives*, *32*(2), 155–172.

Meraner, M., Musshoff, O., & Finger, R. (2018). Using involvement to reduce inconsistencies in risk preference elicitation. *Journal of Behavioral and Experimental Economics*, *73*, 22–33.

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behavior*, *1*, 803–809.

Reynaud, A. & Couture, S. (2012). Stability of risk preference measures: Results from a field experiment on French farmers. *Theory and Decision*, *73*(2), 203–221.

Ruggeri, M. & Coretti, S. (2015). Do probability and certainty equivalent techniques lead to inconsistent results? Evidence from gambles involving life-years and quality of life. *Value in Health*, *18*, 413–420.

Sen, A. (1993). Internal consistency of choice. *Econometrica*, *61*(3), 495–521.

Slovic, P. (1964). Assessment of risk taking behavior. *Psychological Bulletin*, *61*(3), 220.

Slovic, P. (1972a). Information processing, situation specificity, and the generality of risk-taking behavior. *Journal of Personality and Social Psychology*, *22*(1), 128–134.

Slovic, P. (1972b). Psychological study of human judgment: Implications for investment decision making. *Journal of Finance*, *27*(4), 779–799.

Smith, A. (1968). *The money game*. New York: Random House.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, *38*(2), 332–382.

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

Sugden, R. (1991). Rational choice: A survey of contributions from economics and philosophy. *The Economic Journal*, *101*(407), 751–785.

Tanaka, T., Camerer, C. F., & Nguyen, Q. (2010). Risk and time preferences: Linking experimental and household survey data from Vietnam. *America Economic Review*, *100*(1), 557–571.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking and Reasoning*, *20*(2), 147–168.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, *95*(3), 371–384.

Tversky, A. & Thaler, R. (1990). Preference reversals. *Journal of Economic Perspectives*, *4*(2), 201–211.

Vosgerau, J. & Peer, E. (2018). Extreme malleability of preferences: Absolute preference sign changes under uncertainty. *Journal of Behavioral Decision Making*, *32*, 38–46.

Wakker, P. P. (2008). Explaining the characteristics of the power (CRRA) utility family. *Health Economics*, *17*, 1329–1344.

Wakker, P. P. & Deneffe, D. (1996). Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Science*, *42*(8), 1131–1150.

Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*, 263–290.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making*, *26*, 198–212.

Wilcox, N. T. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. In J. Cox & G. Harrison (Eds.), *Risk aversion in experiments* (pp. 197–292). Research in Experimental Economics 12. Bingley, UK: Emerald.

Zhou, W. & Hey, J. D. (2017). Context matters. *Experimental Economics*, *21*(4), 723–756.

A. Appendix

In order to investigate individual-level characteristics that could potentially explain the observed (in)variance of revealed risk preferences across the four risk preference elicitation tasks, subjects in the experiment were asked to answer additional questionnaires. For each subject, specific questions on the tasks followed the same ordering as the completion of each task to avoid confusion. To simplify discrimination between the four different decision problems, the tasks were labeled with color names and highlighted in the respective color whenever displayed to subjects (see the Electronic Supplementary Material for screenshots of the entire experiment).

A.1. Questionnaires

Immediately after subjects had made their decision in any of the four tasks, they self-reported how risky they perceive their decision to be and how confident they feel with the particular choices they made. Each decision, as participants have just completed it, was depicted on screen and questions were answered on a scale from 1 ("not at all risky/confident") to 7 ("very risky/confident"): *(i)* "How risky do you consider your own decision (indicated above)?" and *(ii)* "How confident do you feel with your decision indicated above?" Experimental results of the answers to these questions are reported in Panel A and B of Fig. S1.

After completing all elicitation methods, subjects answered additional questionnaires explicitly comparing the four tasks. For completing the comparative questionnaires, subjects received a payment of €3.00. Questions were answered on a scale from 1 ("not agree at all") to 7 ("fully agree") and read as follows: *(i)* "the task is easy to understand and can be answered straightforwardly," *(ii)* "the task involves complex calculations and requires deliberating on the trade-off between expected outcomes and the inherent riskiness of the different outcomes," and *(iii)* "completing the task is annoying and boring." Experimental results of the answers to these questions are reported in Panel C, D, and E of Fig. S1, respectively. Tab. S1 reports correlations between the questionnaire items and the preference stability index.

Note several characteristics in subjects' responses to the five questions as depicted in Fig. S1: First, general levels of mean responses on confidence and simplicity of choices across tasks are fairly high, whereas they are rather low for answers on boredom and complexity. In general, this can be considered as good news for experimental research on risk preferences. Second, there is substantial variation in response levels of riskiness, confidence, simplicity and complexity. Thus, subjects seem to perceive the tasks and their choices across methods quite differently. In addition to the findings reported in the paper, this result calls for more caution in choosing a particular method to elicit risk preferences. Third, while reported complexity of the tasks seems to clearly relate to subjects' mistakes, as reported in the paper, self-assessed confidence on a subject's decision does not. We conjecture that the assessed confidence does encompass a variety of subjects' attributes, rather than only relating to complexity and difficulty of methods. Fourth, although we cannot conclusively determine whether a set of different preferences across methods influences subjects' choices in our experiment, this could explain the pattern observed in the self-reported assessment of confidence. For instance, confidence is lowest in the BRET with the minimum possible gain (namely zero, regardless of how many boxes are selected) in our experimental set-up (for a discussion on the impact of the availability of safe options, see, e.g. Crosetto and Filippin, 2017). However, this reasoning is only speculative and further examination of such a relation has to be left for future research.

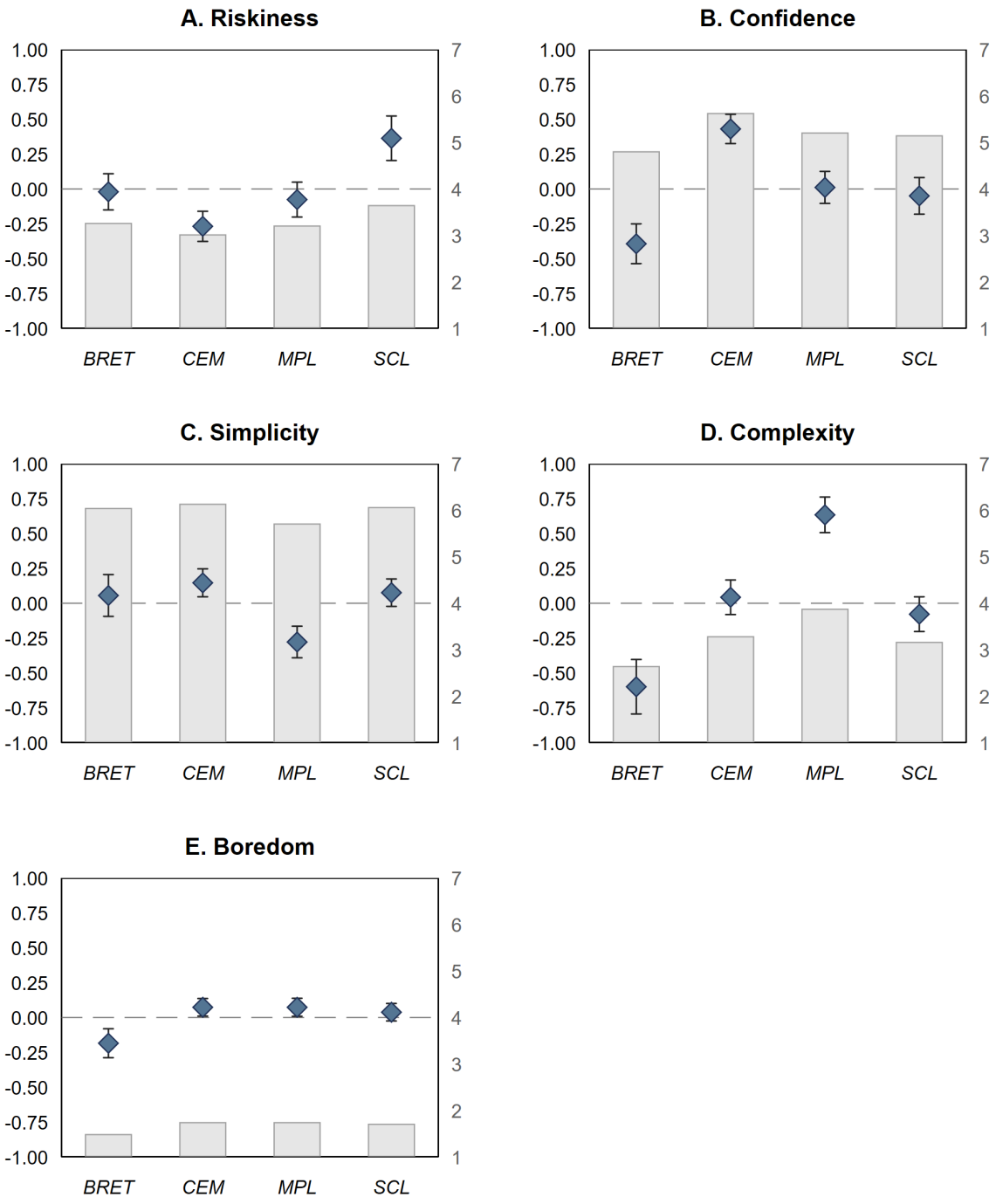

Figure S1: Subject-level demeaned scores (left y-axis) and mean levels (right y-axis) for self-reported answers to survey questions on (A) riskiness of own decision, (B) confidence in own decision, (C) simplicity of task instructions, (D) complexity of calculations involved, and (E) boredom, separated by tasks. In all panels, error bars indicate 95% confidence intervals; $n = 198$. BRET, CEM, MPL, and SCL denote the "bomb" risk elicitation task, the certainty equivalent method, the multiple price list, and the single choice list, respectively.

At the very end of the experiment, participants were asked to state their preferences for a task in future experiments (as a single choice). Of the 198 subjects 30.8% prefer BRET, 31.3% CEM, 21.7% MPL, and 16.2% SCL. There is not much difference between BRET and CEM, which are favoured, while SCL is the least preferred task.

Table S1: Correlations of the preference stability index and responses to the questionnaire items. The lower triangular matrix depicts Spearman rank correlations; the upper triangular matrix reports polychoric correlations. p-values are reported in parentheses ($n = 198$). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. BRET, CEM, MPL, and SCL denote the "bomb" risk elicitation task, the certainty equivalent method, the multiple price list, and the single choice list, respectively.

	Stab.	Risk.	Conf.	Simp.	Comp.	Bore.
Stability Index		0.127***	0.025	−0.050	0.053	0.049
		(0.001)	(0.502)	(0.231)	(0.163)	(0.273)
Riskiness	0.143***		−0.028	−0.015	0.093*	0.080
	(0.000)		(0.464)	(0.713)	(0.014)	(0.072)
Confidence	0.023	−0.025		0.191***	−0.005	−0.076
	(0.516)	(0.481)		(0.000)	(0.906)	(0.086)
Simplicity	−0.048	−0.038	0.156***		−0.394***	−0.177***
	(0.179)	(0.281)	(0.000)		(0.000)	(0.000)
Complexity	0.076*	0.099**	−0.008	−0.358***		0.127**
	(0.032)	(0.006)	(0.829)	(0.000)		(0.004)
Boredom	0.032	0.078*	−0.045	−0.129***	0.107**	
	(0.372)	(0.028)	(0.206)	(0.000)	(0.003)	

A.2. Domain Attribution

In comparing the four elicitation methods, subjects were also asked whether they associate the decision problem with an investment, gambling, or insurance domain using a drop-down field with the three possible options. Responses per task are reported in Table S2. Preference stability rate for pairwise comparisons of risk preference elicitation methods separated by attributions to the same domain or different domains is reported in S2.

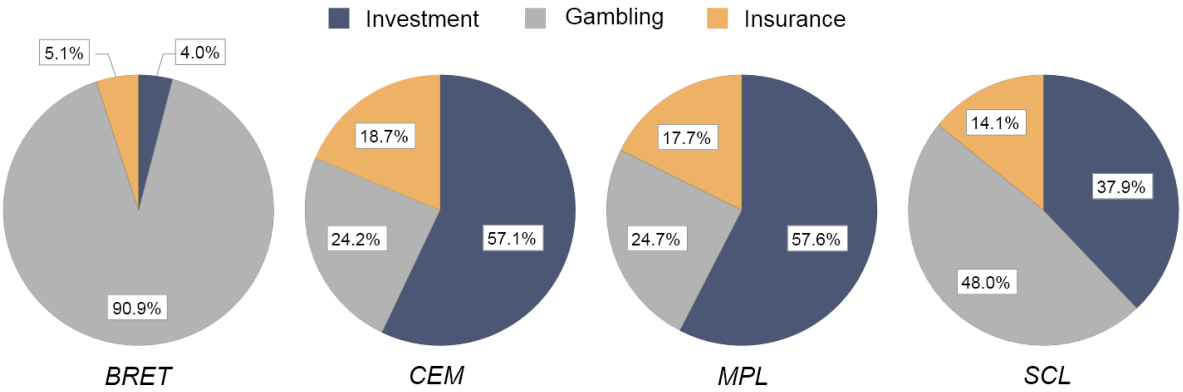

Figure S2: Subjects' attribution of domains—(i) investment, (ii) gambling, and (iii) insurance—separated by risk preference elicitation methods. $n = 198$. BRET, CEM, MPL, and SCL denote the "bomb" risk elicitation task, the certainty equivalent method, the multiple price list, and the single choice list, respectively.

Table S2: Preference stability rate for pairwise comparisons of risk preference elicitation methods separated by whether the tasks are perceived to belong to the same domain or to different domains. Number of observations per class are reported in parentheses. Test statistics and p-values of $\chi^2(1)$-tests on differences between "same domain" and "different domain" are depicted in the lower panel.

Domain ...	BRET–CEM	BRET–MPL	BRET–SCL	CEM–MPL	CEM–SCL	MPL–SCL
same	50.0% (42)	21.7% (46)	35.8% (81)	54.2% (118)	70.8% (89)	50.0% (86)
different	50.4% (143)	20.9% (139)	33.7% (104)	68.7% (67)	68.8% (96)	48.5% (99)
$\chi^2(1)$	0.002	0.016	0.093	3.686	0.091	0.042
p-value	0.968	0.900	0.761	0.055	0.763	0.837

A.3. Task Comprehension and Numeracy

Subjects were asked to estimate (i) the expected payoff, (ii) the probability to earn less than €5.50 and (iii) the probability to earn more than €14.50 for the risk neutral decision (depicted as a screenshot) in each of the four tasks. On average, subjects responses deviated from the correct answers by 164.4% (sd = 92.4%) in BRET, 111.7% (sd = 69.7%) in CEM, 177.0% (sd = 95.6%) in MPL, and 57.7% (sd = 60.9%) in SCL.

In addition, we included an 8-item Rasch-validated numeracy inventory (Weller et al., 2013), including two items on cognitive reflection[19], to assess participants' numerical skills. The numeracy inventory was incentivized with €0.50 for each correct answer. On average, participants correctly answered 5.49 ($sd = 1.57$) out of 8 questions.

While subjects have rather high levels of numeracy, their estimation of expected returns and probabilities show strong deviations from the correct answers. Furthermore, it is noteworthy that actual errors in estimations are not necessarily in line with the perceived complexity of tasks. While subjects seem to be able to assess the susceptibility to errors in making choices in the tasks according to their risk preferences, as argued in the paper, this self-assessment seems not to be directly related to the ability to calculate expected returns and probabilities. This further corroborates the—often implicitly made—assumption that subjects can reveal their preferences in the tasks without explicitly being able to correctly solve the calculations behind the tasks' lotteries.

A.4. Order Effects

In order to prevent that the ordering of risk preference elicitation procedures systematically affects subjects' risk-taking behavior, the sequence of tasks in the experiment was randomized on the subject level. Yet, despite the randomization, one could hypothesize that risk preferences, and/or subjects' susceptibility to making mistakes in evaluating the alternatives, might be affected by the task ordering. For instance, it might be the case that subjects try to balance the overall risk they take in the experiment, i.e., subjects may take systematically more (less) risk if their decision in the previous task involves a low (high) level of risk. Likewise, due to learning effects, subjects might be less prone to making errors in evaluating the choices in tasks that appear towards the end of the sequence; or, on the contrary, one could argue that fatigue increases the likelihood of making mistakes, etc.

[19] The inventory proposed by Weller et al. (2013) includes two of the three cognitive reflection test items introduced by Frederick (2005). As these questions have been used repeatedly in our laboratory and correct answers to the questions might be known, we replaced these items by two questions proposed by Toplak et al. (2014).

Table S3: (A) Maximum likelihood estimates of structural models with Fechner error terms for each of the four positions in the random task sequence. Standard errors, clustered on the subject level, are reported in parentheses. **(B)** Pairwise differences in point estimates of risk preference parameters φ (lower-triangular matrix) and the standard deviation of noise parameters σ (upper-triangular matrix) between the four positions in the task sequence. p-values are based on pairwise Wald tests. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Panel A	Order = 1	Order = 2	Order = 3	Order = 4
φ	0.601***	0.645***	0.541***	0.548***
	(0.057)	(0.059)	(0.054)	(0.050)
σ	0.346***	0.253***	0.365***	0.348***
	(0.064)	(0.046)	(0.072)	(0.066)
$\ln L$	−2,753	−4,222	−2,804	−3,250
No. of Obs.	5,306	7,798	5,346	6,102
Clusters	198	198	198	198

Panel B	Order = 1	Order = 2	Order = 3	Order = 4
Order = 1		0.093	−0.019	−0.001
Order = 2	−0.044		−0.112	−0.094
Order = 3	0.060	0.104		0.018
Order = 4	0.053	0.097	−0.007	

To rule out that spurious effects drive the results reported in the main text, Tables S3 and S4 summarize additional analyses examining potential order effects. In particular, Table S3A reports maximum likelihood estimates of the structural model (as described in Section 4) for each of the four positions in the random task sequence. Table S3B shows the pairwise differences in point estimates of φ and σ between the four positions in the sequence. Apparently, none of the differences—neither for the CRRA coefficient φ nor for the standard deviation of the noise parameter σ—is statistically significantly different from zero.

To rule out that our main findings are impaired by some systematic effect of a particular task on the succeeding one, we estimate the structural model for each of the four risk preference elicitation methods, controlling for the preceding one in the random sequence. Indeed, as depicted in Table S4, none of the dichotomous controls—neither for φ nor σ—turns out to be statistically significant, suggesting that our results are not affected by potential interrelations between preceding and succeeding tasks in the ordering.

The analyses above only provide insights for potential effects of the task ordering on the estimates of the mean parameters in the structural model, but not on individual-level instability of preference estimates across methods. To address the latter, we examine potential order effects with respect to the stability index defined in Section 4. Given that there are four tasks, each of which might take any of the four positions in the sequence, there are $n = 4!$ possible permutations.[20] To evaluate whether specific permutations induce significantly different stability indices, we conduct 24 t-tests, each comparing the mean stability index of one particular permutation to the mean stability index of the remaining 23 task

[20] For the randomization in our experiment it turns out that each of the 24 possible permutations was realized as the task ordering for at least two subjects, whereas the maximum number of subjects who faced the same permutation was 19.

Table S4: Maximum likelihood estimates of structural models with Fechner error terms for each of the four risk preference elicitation methods, controlling for the preceding task. BRET, CEM, MPL, and SCL denote the "bomb" risk elicitation task, the certainty equivalent method, the multiple price list, and the single choice list, respectively. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	BRET	CEM	MPL	SCL
φ (Constant)	0.680***	0.967***	0.615***	0.321***
	(0.042)	(0.133)	(0.070)	(0.071)
Prev. Task = BRET		−0.283	0.013	0.077
		(0.226)	(0.093)	(0.093)
Prev. Task = CEM	−0.036		−0.033	0.058
	(0.056)		(0.090)	(0.095)
Prev. Task = MPL	−0.078	−0.333		0.128
	(0.061)	(0.237)		(0.104)
Prev. Task = SCL	−0.111	0.006	−0.028	
	(0.060)	(0.273)	(0.101)	
σ (Constant)	0.044***	0.184***	0.954***	0.707***
	(0.004)	(0.051)	(0.126)	(0.110)
Prev. Task = BRET		0.151	−0.087	−0.065
		(0.127)	(0.160)	(0.150)
Prev. Task = CEM	0.005		−0.172	0.030
	(0.007)		(0.177)	(0.153)
Prev. Task = MPL	−0.002	0.218		0.095
	(0.005)	(0.168)		(0.182)
Prev. Task = SCL	−0.001	0.048	0.243	
	(0.006)	(0.125)	(0.201)	
$\ln L$	−5,235	−452	−592	−571
No. of Obs.	19,800	1,782	1,980	990
Clusters	198	198	198	198

Table S5: (A) Maximum likelihood estimates of random parameter models for each of the four risk preference elicitation methods. Standard errors, clustered on the subject level, are reported in parentheses. **(B)** Pairwise differences in point estimates of risk preference parameters φ (lower-triangular matrix) and precision parameters λ (upper-triangular matrix) between the four risk preference elicitation methods. p-values are based on pairwise Wald tests. BRET, CEM, MPL, and SCL denote the "bomb" risk elicitation task, the certainty equivalent method, the multiple price list, and the single choice list, respectively. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Panel A	BRET	CEM	MPL	SCL
φ	0.653***	0.896***	0.696***	0.579***
	(0.023)	(0.004)	(0.012)	(0.014)
$ln\lambda$	3.733***	2.999***	4.056***	4.837***
	(0.174)	(0.851)	(0.052)	(0.650)
$\ln L$	−12,153	−838	−1,207	−536
No. of Obs.	19,404	1,386	1,980	990
Clusters	198	198	198	198

Panel B	BRET	CEM	MPL	SCL
BRET		0.779***	−0.320	−0.947***
CEM	0.241***		−1.099***	−1.726***
MPL	0.043	−0.199***		−0.627*
SCL	−0.078*	−0.320***	−0.121*	

sequences. It turns out that the stability indices do not statistically differ from other permutations for any of the 24 different sequences (t-values vary between 0.033 and 1.754; corresponding p-value range from 0.973 to 0.081, respectively).

Overall, the analyses summarized above provide strong evidence that the results presented in the main text are not spurious in the sense that they might be the result of order or learning effects.

A.5. Alternative Model Specification

In addition to the estimates based on the random utility model reported in the main text, we report estimates of the random parameter model in Table S5. It is reassuring that our main findings are corroborated by qualitatively similar results: while the point estimates of the CRRA parameter φ turn out to be higher for all tasks (which is in line with the results reported by Apesteguia and Ballester, 2018), the ordering of parameter estimates across tasks is preserved and the patterns of statistically significant differences between tasks remain similar.

In a recent article, Apesteguia and Ballester (2018) prove that random utility models may violate monotonicity in the sense that the probability of choosing the more risky alternative in a binary choice setting is an increasing function of the risk preference parameter. Yet, Apesteguia and Ballester (2018) show that random parameter models, as introduced by Eliashberg and Hauser (1985) and Loomes and Sugden (1995), are always monotone in the choice probability.

In contrast to binary random utility models, where the noise term is modelled to distort the evaluation of expected utilities of the two alternatives, the noise term in random parameter models distorts the decision maker's preference parameter. That is, the decision maker is assumed to choose the alternative that maximizes the utility given a particular coefficient of risk aversion φ, distorted by a common random error ϵ. In the random parameter model, the probability of choosing alternative B has the closed form of $e^{\lambda\varphi^*}/(e^{\lambda\varphi^*} + e^{\lambda\varphi})$, where φ^* refers to the CRRA parameter which equates the expected utilities of the two lotteries, i.e., $u_{\varphi^*}(A) = u_{\varphi^*}(B)$, and λ denotes a precision parameter which is inversely related to the variance of a random noise term (Apesteguia and Ballester, 2018). Note that a decrease in φ (i.e., a decrease in risk aversion) implies a decrease in the denominator, guaranteeing that the choice probability increases and monotonicity is preserved.

Despite the advantage of the random parameter model, we chose to report results in the Appendix for two reasons: First, the value of φ^* is practically not determinable for the dominated choices in the BRET and CEM i.e. for 2 out of 9 binary choices in the CEM and for 2 out of 99 binary choices in the BRET. Applying the random parameter model, thus, implies that 396 observations need to be dropped from the analysis for both the BRET and the CEM. Second, the solutions for the values of φ^* turn out to be labile for several binary choices, in particular for very low ($k < 5$) and very high ($k > 90$) numbers of selected boxes in the BRET, but also for the most and least risk averse decisions in the other three tasks. Given the properties of the power function representing subjects' utility, solutions to the non-linear equations are overly sensitive to marginal deviations and, thus, computationally hard to approximate. The extent to which the parameter estimates of a particular task are impaired by these effects may well be systematic in nature since the parameter ranges covered by the different elicitation procedures – and thereby the values of the solutions for φ^* and, as a result, the precision of solutions for φ^* — vary substantially. Irrespective of which model is considered superior given our data, it is eventually reassuring that our main results are robust in both specifications.

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

https://www.uibk.ac.at/eeecon/wopec/

2019-19 **Felix Holzmeister, Matthias Stefan:** The risk elicitation puzzle revisited: Across-methods (in)consistency?

2019-18 **Katharina Momsen, Markus Ohndorf:** Information Avoidance, Selective Exposure, and Fake(?) News-A Green Market Experiment

2019-17 **Stjepan Srhoj, Bruno Skrinjaric, Sonja Radas, Janette Walde:** Cognitive Skills and Economic Preferences in the Fund Industry

2019-16 **Adam Farago, Martin Holmen, Felix Holzmeister, Michael Kirchler, Michael Razen:** Closing the Finance Gap by Nudging: Impact Assessment of Public Grants for Women Entrepreneurs

2019-15 **Christopher Kah, Daniel Neururer:** Generiert der stationäre Buchhandel positive Nachfrageeffekte und verhilft dadurch dem Kulturgut Buch bei seiner Verbreitung? - Ein natürliches Experiment

2019-14 **Stjepan Srhoj, Michael Lapinski, Janette Walde:** Size matters? Impact evaluation of business development grants on SME performance

2019-13 **Andrea M. Leiter, Engelbert Theurl:** DETERMINANTS OF PREPAID SYSTEMS OF HEALTH-CARE FINANCING - A WORLDWIDE COUNTRY-LEVEL PERSPECTIVE

2019-12 **Michael Razen, Michael Kirchler, Utz Weitzel:** Domain-Specific Risk-Taking Among Finance Professionals

2019-11 **Jonathan Hall, Rudolf Kerschbamer, Daniel Neururer, Eric Skoog:** Uncovering sophisticated discrimination with the help of credence goods markups - evidence from a natural field experiment

2019-10 **Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** Collective intertemporal decisions and heterogeneity in groups

2019-09 **Morten Hedegaard, Rudolf Kerschbamer, Daniel Müller, Jean-Robert Tyran:** Distributional Preferences Explain Individual Behavior Across Games and Time

2019-08 **Daniel Müller, Sander Renes:** Fairness Views and Political Preferences - Evidence from a representative sample

2019-07 **Florian Lindner, Michael Kirchler, Stephanie Rosenkranz, Utze Weitzel:** Social Status and Risk-Taking in Investment Decisions

2019-06 **Christoph Huber, Julia Rose:** Individual attitudes and market dynamics towards imprecision

2019-05 **Felix Holzmeister, Jürgen Huber, Michael Kirchler, Florian Lindner, Utz Weitzel, Stefan Zeisberger:** What Drives Risk Perception? A Global Survey with Financial Professionals and Lay People

2019-04 **David M. McEvoy, Tobias Haller, Esther Blanco:** The Role of Non-Binding Pledges in Social Dilemmas with Mitigation and Adaptation

2019-03 **Katharina Momsen, Markus Ohndorf:** When do people exploit moral wiggle room? An experimental analysis in a market setup

2019-02 **Rudolf Kerschbamer, Daniel Neururer, Matthias Sutter:** Credence goods markets and the informational value of new media: A natural field experiment

2019-01 **Martin Geiger, Eric Mayer, Johann Scharler:** Inequality and the Business Cycle: Evidence from U.S. survey data

2018-18 **Matthias Sutter, Jürgen Huber, Michael Kirchler, Matthias Stefan, Markus Walzl:** Where to look for the morals in markets?

2018-17 **Rene Schwaiger, Michael Kirchler, Florian Lindner, Utz Weitzel:** Determinants of investor expectations and satisfaction. A study with financial professionals

2018-16 **Andreas Groll, Julien Hambuckers, Thomas Kneib, Nikolaus Umlauf:** LASSO-Type Penalization in the Framework of Generalized Additive Models for Location, Scale and Shape

2018-15 **Christoph Huber, Jürgen Huber:** Scale matters: Risk perception, return expectations, and investment propensity under different scalings

2018-14 **Thorsten Simon, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis:** Lightning prediction using model output statistics

2018-13 **Martin Geiger, Johann Scharler:** How do consumers interpret the macroeconomic effects of oil price fluctuations? Evidence from U.S. survey data

2018-12 **Martin Geiger, Johann Scharler:** How do people interpret macroeconomic shocks? Evidence from U.S. survey data

2018-11 **Sebastian J. Dietz, Philipp Kneringer, Georg J. Mayr, Achim Zeileis:** Low visibility forecasts for different flight planning horizons using tree-based boosting models

2018-10 **Michael Pfaffermayr:** Trade creation and trade diversion of regional trade agreements revisited: A constrained panel pseudo-maximum likelihood approach

2018-09 **Achim Zeileis, Christoph Leitner, Kurt Hornik:** Probabilistic forecasts for the 2018 FIFA World Cup based on the bookmaker consensus model

2018-08 **Lisa Schlosser, Torsten Hothorn, Reto Stauffer, Achim Zeileis:** Distributional regression forests for probabilistic precipitation forecasting in complex terrain

2018-07 **Michael Kirchler, Florian Lindner, Utz Weitzel:** Delegated decision making and social competition in the finance industry

2018-06 **Manuel Gebetsberger, Reto Stauffer, Georg J. Mayr, Achim Zeileis:** Skewed logistic distribution for statistical temperature post-processing in mountainous areas

2018-05 **Reto Stauffer, Georg J. Mayr, Jakob W. Messner, Achim Zeileis:** Hourly probabilistic snow forecasts over complex terrain: A hybrid ensemble postprocessing approach

2018-04 **Utz Weitzel, Christoph Huber, Florian Lindner, Jürgen Huber, Julia Rose, Michael Kirchler:** Bubbles and financial professionals

2018-03 **Carolin Strobl, Julia Kopf, Raphael Hartmann, Achim Zeileis:** Anchor point selection: An approach for anchoring without anchor items

2018-02 **Michael Greinecker, Christopher Kah:** Pairwise stable matching in large economies

2018-01 **Max Breitenlechner, Johann Scharler:** How does monetary policy influence bank lending? Evidence from the market for banks' wholesale funding

University of Innsbruck

Working Papers in Economics and Statistics

2019-19

Felix Holzmeister, Matthias Stefan

The risk elicitation puzzle revisited: Across-methods (in)consistency?

Abstract
With the rise of experimental research in the social sciences, numerous methods to elicit and classify people's risk attitudes in the laboratory have evolved. However, evidence suggests that people's attitudes towards risk may change considerably when measured with different methods. Based on a with-subject experimental design using four widespread risk preference elicitation methods, we find that different procedures indeed give rise to considerably varying estimates of individual and aggregate level risk preferences. Conducting simulation exercises to obtain benchmarks for subjects' behavior, we find that the observed heterogeneity in risk preference estimates across methods looks qualitatively similar to the heterogeneity arising from independent random draws from choices in the experimental tasks, despite significantly positive correlations between tasks. Our study, however, provides evidence that subjects are surprisingly well aware of the variation in the riskiness of their choices. We argue that this calls into question the common interpretation of variation in revealed risk preferences as being inconsistent.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)