Universität Basel

Fakultät für Psychologie

Probing the Determinants of Risky Decision Making in the Wild

Inauguraldissertation zur Erlangung der Würde eines Doktors der Philosophie vorgelegt der Fakultät für Psychologie der Universität Basel von

Oliver Tom Schürmann

aus Basel, Schweiz

Basel, 2017

Universität Basel

Fakultät für Psychologie

Genehmigt von der Fakultät für Psychologie auf Antrag von

Prof. Dr. Jörg Rieskamp

Prof. Dr. Ralph Hertwig

Datum des Doktoratsexamen: 09.10.2017

Prof. Dr. Roselind Lieb

Probing the Determinants of Risky Decision Making in the Wild

Oliver Schürmann

University of Basel

Author Note

Brief framework for the cumulative dissertation based on:

Schürmann, O., Andrascewicz, S., & Rieskamp, J. (2017)

Schürmann, O., Frey, R., Hertwig, R., Rieskamp, J., & Pedroni, A. (2017)

Schürmann, O., Pleskac, T.J., & Frey, R. (2017)

Probing the Determinants of Risky Decision Making in the Wild

Declaration

I, Oliver Tom Schürmann, born on January 14th, 1987 in Basel, Switzerland, hereby declare the following:

(i) My cumulative dissertation is based on three manuscripts, all of which are submitted for publication. I have been primary responsible for the ideas, data collection and writing of the manuscripts. This characterization of my contribution is in agreement with my co-authors' views.

(ii) I only used the resources indicated.

(iii) I marked all the citations.

Basel, September 21st, 2017

Oliver Schürmann

Acknowledgments

This work is not so much the achievement of myself as it is the combined effort of my co-authors, advisors, colleagues, my family and friends, without whom this work would have never even begun.

I thank my co-authors and advisors; Andreas Perdoni, who embarked on this endeavor with me and always reminded me of why we do what we do; Tim Pleskac, who taught me not only to model my data and commit to strict ethical research practices, but also encouraged me time and again to "keep pushing"; Sandra Andraszewicz for coping with me and for brightening up the world of economic risk taking; and Renato Frey, who always gave quick and on-point advice. I would like to thank my primary advisors Jörg Rieskamp for providing me the opportunity to work in his group and for teaching me to stick to the details as well as Ralph Hertwig, for allowing me to be part of his team in Berlin, learn from his perspicacity and for getting me hooked on decision making in the first place. I would also like to thank Laura Wiles and Susannah Goss for editing my manuscripts.

I am deeply in debt to my colleagues at the Center for Economic Psychology in Basel as well the Center for Adaptive Rationality in Berlin, for providing me the best work-environment and hands-on ruses of Ph.D. life I could have wished for. I owe special thanks to my fellow doctoral students in Basel (Michael, Janine, Laura, Sebastian, Rebecca and Regina) and Markus, for down to earth discussions and perspectives.

By all the writing skills provided to me in my time as a doctoral student, I cannot express how much this thesis depended on my family and friends. It is thanks to their immense amount of dedication, understanding, and care, this work found to an end. It is them who I dedicate this work to, for without their contribution to my life; I would not be writing anything...ever. Finally, I dedicate this work to Alois and Waltraud, who saw the beginning of this work but will never get to see the end of it.

Abstract

Risky decision making carries many of our behaviors in everyday life. Behavioral researchers have been perpetually probing risky decision making using a plethora of different measures and since over two hundred years, different models have been integrating the results of these measures. Despite this effort, it remains elusive how determinants of risk taking, probed in the enclosed space of the behavioral laboratory, map to actual everyday decisions in the wild. What is the structure of these determinants? Do people have stable *preferences* for risky decision making over various domains? And how is the *perception* of risk influencing the process of decision making?

This dissertation focuses on two major determinants of risky decision making; *risk preferences*, and *risk perceptions*. The first part includes two manuscripts focusing on quantifying risk preferences that are relevant in real-life. The second part of the dissertation investigates risk perceptions, which come into action especially when people do not have full information about decision-options, which defines most everyday-behavior. Manuscript one probes different existing and newly developed risk-preference measures in terms of retest-reliability and validity. The new measure builds advantages of various existing measures and proves to be more reliable than existing measures of its kind. In manuscript two, we provide an example of how specific and general risk-preference measures boost the predictive power of risky decision making in actual street crossing behavior. Finally, in part two, manuscript three quantifies risk perceptions in a sequential risk task and shows that when knowledge about decision options is incomplete, people rely strongly on their subjective probabilities to make decisions in the task. In turn, these subjective probabilities are very vulnerable to initial experience.

Harnessing decades of research on risky decision making, the results of this dissertation build on over 50 different risk measures to target the gap between decision-theory and everyday behavior. Results from these measures are evaluated using mathematical models of decision making, to make assumptions about the cognitive determinants of risky decision making in the wild.

General Introduction

> *... it seems clear that all men cannot use the same rule to evaluate the [risk]... thus there is no doubt that a gain of a thousand ducats is more significant to a pauper than to a rich man though both gain the same amount.*
>
> Daniel Bernoulli, Basel, 1738

Over 200 years ago, Daniel Bernoulli wrote his seminal work on the quantification of risk-taking behavior, which became a milestone in theory development of decision making (Bernoulli, 1738/1954). With his elaborations about the utility of a choice, he developed an extension of the until then dominating theory of expected value (EV; cf. Ore, 1960). The EV of a risky choice option (i.e., an option that yields at least two outcomes with non-zero probabilities; the EV of a sure option is equal to its outcome magnitude) can be obtained by taking the sum of the product of each possible outcome and its corresponding probability. According to normative decision making theories, rational agents should always maximize EV (i.e., always choose the option with the largest EV). For example, consider a simple game with two options A and B: A is a lottery that yields 2000 Swiss francs (SFR) with a probability of .5, and otherwise nothing, while B offers a sure gain of 900 SFR (probability of 1). A has an EV of 1000 SFR ($.5 \times 2000$ SFR) and B an EV of 900 SFR (1×900 SFR). Based on the EV theory, everyone, no matter their personal state and disregarding characteristics of the environment, should choose the lottery A. However, Bernoulli observed that this is neither what people usually do nor what they *should* do. As he stated:

> "Somehow a very poor fellow obtains a lottery ticket that will yield with equal probability either nothing or twenty thousand ducats. Will this man evaluate his chance of winning at ten thousand ducats? Would he not be ill-advised to sell this lottery ticket for nine thousand ducats? To me it seems that the answer is in the negative. On the other hand I am inclined to believe that a rich man would be ill-advised to refuse to buy the lottery

> ticket for nine thousand ducats. If I am not wrong then it seems clear that all men cannot use the same rule to evaluate the gamble."

(Bernoulli, 1738/1954, , p. 24)

Bernoulli argued that people's preferences for risky endeavors are (and should be) inter-individually different, depending on their personal characteristics (he mainly referred to their a-priori wealth). In that way, "utility" replaces the plain EV or "price" of an option. Therefore, Bernoulli's theory is often referred to as expected utility (EU) theory. Today, researchers build on these theoretical structures to quantify risk-preferences in every decision.

Until today, however, it remains elusive how risk preferences influence everyday, real-life decisions. What is the cognitive structure of risk-preferences? How are risk-preferences integrated into our cognitive system of decision making? Is there a general factor of risk-preference or do people build risk-preferences "online" by integrating cues of the environment? How do people perceive risk when the criteria of options are not described but have to be estimated?

To follow up on these questions, this dissertation harnesses the investigative power of two methodologies: first, probing (i.e., measuring) risk-taking and risk-preferences using a variety of measures in different real-life situations and second, integrating these measurements into different frameworks of decision making (i.e., through mathematical cognitive modeling). In doing so, my dissertation titled *Probing the Determinants of Risky Decision Making in the Wild* aims to bridge the gap between laboratory behavior and decision making in everyday life, focusing on determinants that influence the connection between decision-theoretical assumptions and real-world behavior.

After a brief theoretical introduction, this framework consists of two parts. The first part is concerned with the measurement and structure of people's risk-preferences and how, in turn, preferences determine decisions in everyday life. In manuscript one (Schürmann, Andrascewicz, & Rieskamp, 2017), we discuss and test a number of risk-preference measures that aim to quantify individual preference parameters. We identify some important attributes of successful risk measures such as incorporating

losses and framing the questions as real-life contexts. We introduce a new measure, the Basel risk preference measure (BAR) that incorporates these important attributes to increase the reliability and validity of risk-preference measurements.

In manuscript two (Schürmann, Frey, Hertwig, Rieskamp, & Pedroni, 2017), we address the specificity and structure of risk-preferences. While specific situations arguably trigger specific preferences, it has been recently shown that risk-preferences likely also include a general factor(Frey, Pedroni, Mata, Rieskamp, & Hertwig, 2017; Highhouse, Nye, Zhang, & Rada, 2016; Menkhoff & Sakha, 2017). By incorporating both facets of risk-preference with domain-specific and domain-general measures of risk-preference, we report the increase of predictive power of naturalistic street-crossing behavior after a delay of ten months.

Part two of this dissertation, on the other hand, focuses on the fact that in everyday life, people often have limited knowledge about the available decision options. While in Bernoulli's lottery example the probabilities and payoffs are clearly specified, in everyday life, exact probabilities are often unknown or unknowable (Hertwig & Erev, 2009; Knight, 1921; Simon, 1955). Manuscript three (Schürmann, Pleskac, & Frey, 2017) thus focuses on the perception of probabilities, their influence on risky decision making and how they adapt to initial experiences.

Mind the Gaps: Terminologies of Risk

When investigating determinants of risky decision making in the wild, it is crucial first to consider the different terminologies of risk that have been shown to influence the connection between laboratory and everyday decisions. While in the literature many terminologies of risky behavior are being used interchangeably (e.g., risk-taking, risky behavior, risky decision making), two major gaps of extant terminologies should be given special attention.

First, economists typically define risk regarding the variability of outcomes, while psychologists and laypeople view risk as the experience of possible loss or harm (March & Shapira, 1987; Schonberg, Fox, & Poldrack, 2011). While in some cases both

definitions come to similar conclusions, the difference needs to be taken into account when measuring risk-preferences to predict behavior in everyday life or when comparing different risk-measures with each other (Frey et al., 2017; Pedroni et al., 2017). Manuscript one approaches this differentiation and exemplifies the crucial inclusion of losses in lottery-based risk-preference measurements.

The second gap in terminologies refers to the differentiation between decisions made under full disclosure of all information (referred to as decisions from description), or decision made when full information is unavailable and has to be acquired through experience (cf. decisions from experience; Hertwig, Barron, Weber, & Erev, n.d.; Hertwig & Erev, 2009). While economists typically use the terms decisions under risk (from description) in contrast to decisions under uncertainty or ambiguity (e.g., from experience; Knight, 1921), in the context of this framework, and akin to most psychologists' definitions, I use the term risk not only for cases where all information is fully revealed. However, the distinction between these two cases is a crucial one and directly maps to some of the differences between decision making in the lab and everyday life. We speak to decisions from experience and how the perception of unknown (or unknowable) properties of decision options influence risky behavior in manuscript three of this dissertation.

Measures of Risky Decision Making

Arguably, decision making is influenced by many different aspects of everyday life. Therefore, a crucial question is how to probe risk-taking behavior (and its components; Appelt, Milch, Handgraaf, & Weber, 2011). Over the course of the last couple decades, a plethora of different risk-taking measures have been developed and tested. To date, two major procedures can be distinguished: behavioral measures and self-report measures (cf. Frey et al., 2017). While behavioral measures infer risk-preferences from choices in lotteries and other games, self-report measures rely on people's introspection by asking them questions about past behavior or behavior planned for the future.

Behavioral measures of risk taking infer risk-preferences from (often incentivized)

observed behavior in controlled experimental tasks. Typically, these tasks ask participants to make choices among well-defined monetary lotteries such as the one presented by Bernoulli (see Pedroni et al., 2017). Other behavioral measures use a more naturalistic approach and "pack" the decisions into game-like tasks, often engendering an affective response through exhilaration and tension (see Schonberg et al., 2011, , for a review). One such example is the balloon analogue risk task (BART) in which individuals inflate a virtual balloon, earning points if it does not explode (Lejuez et al., 2002). The number of pumps made in the BART (adjusted BART score) is then used as a measure of risk-preference. Previously, this measure has been found to correlate with a number of real-life risk-taking behaviors (see Lauriola, Panno, Levin, & Lejuez, 2013, for an overview). The clear, quantified structure of behavioral measures allows researchers to capture specific cognitive components in the process of decision making, such as the influence of gains and losses (Köbberling & Wakker, 2005) or learning strategies (Pleskac, 2008).

Self-report measures of risk taking, on the other hand, depend on introspective abilities of individuals (Frey et al., 2017). Typically, these measures ask questions about previously experienced real-life or hypothetical scenarios dealing with certain kinds of risky behaviors. People then rate their willingness to engage in these specific behaviors. For example, the sensation-seeking personality scale presents a number of statement pairs (Zuckerman, Eysenck, & Eysenck, 1978). Within each such pair, one question represents a risky (or "sensation-seeking") option (e.g., "I sometimes like to do things that are a little frightening") and the other one a less risky (or less sensation-seeking) option (e.g., "A sensible person avoids activities that are dangerous"). The sum of all sensation seeking statements chosen then represents the risk preference score.
Self-report measures are often used because they can be implemented quickly and easily Dohmen et al. (2011). Also, they have been found to have good reliability and convergent validity, which behavioral measures sometimes lack(Dohmen et al., 2011; Frey et al., 2017; Pedroni et al., 2017; Weber, Blais, & Betz, 2002). Thus, self-report measures are especially well-suited for predicting real-life risk-taking behavior. In

comparison to behavioral measures, self-report measures are typically limited in their ability to identify strong mechanistic explanations of behavior. To integrate the advantages of both approaches, the conclusions of this dissertation are drawn based on both types of risk-taking measures.

Models of Risky Decision Making

Traditional economic decision making models of risk identify a decision as an implementation of the two decision properties; probability and the values of each possible outcome (see Ore, 1960). Such models discuss how these properties are subjectively interpreted by a decision maker (DM) and the respective scholars' experimental results have shown that DMs' choices often deviate from economic rationality (Ore, 1960; Von Neumann & Morgenstern, 2007). Advancements of these early theories implemented more sophisticated principles to account for additional deviations from economic rationality that have been observed in the mean-time (making them descriptive theories rather than normative theories Kahneman & Tversky, 1979). The arguably most prominent example of these utility-theory derived models is cumulative prospect theory (CPT; Tversky & Kahneman, 1992). The main assumptions that CPT makes are (a) that "people evaluate the outcomes of risky options as gains and losses resulting from a comparison with a context-dependent reference point" (Rieskamp, 2008b, , p.1446). Further, it assumes that people weigh losses larger compared to gains, so that a loss is worse than the same gains is good ("losses loom larger than corresponding gains", Tversky & Kahneman, 1991, p. 1039). Finally, CPT assumes that probabilities are weighted non-linearly (in stark contrast to EV and EU).

Alternative explanations for deviations from economically rational decisions offered the concept of "bounded rationality", and with it heuristic theories of decision making (cf. Gigerenzer, 2006; Simon, 1955) . This line of research argues that in complex decision tasks (such as those people face in the wild), humans employ simple rules of thumb (thus ignoring much information), rather than performing computationally often demanding optimization of the problem at hand. These models

thus focus on the psychological and ecologically adaptive rationality, arguing that, in the wild, people neither have the cognitive capacity nor sufficient knowledge about the decision options to behave in an economical rational sense (Simon, 1955). In line with this is the distinction between decisions from description, on which most economic models are based and decisions from experience, which can often be better explained by models of bounded rationality (Hertwig & Erev, 2009).

Finally, in recent years, the focus of decision sciences has been drawn to determinants that describe the temporal *process* of risky decisions, rather than only the outcome as assumed by "black-box theories" (Busemeyer & Townsend, 1993). A special case of a process model is the Bayesian sequential sampling model (BSR; c.f. Pleskac, 2008; Wallsten, Pleskac, & Lejuez, 2005). The BSR investigates cognitive underpinnings involved in the BART and combines aspects of CPT with a Bayesian learning rule. This model is an example of how learning influences risky decision making. Other process models investigate the decision process as a more detailed information-integration process (Ratcliff & Smith, 2015). In these models, information, or evidence, is accumulated until a decision threshold is reached and the associated decision is made. On example of such a model is the diffusion model.

Across all manuscripts included in this dissertation, we used different models of cognition and compared them with each other to adequately implement and interpret the findings of the different risk-measures we used. As has been shown in a whole body of literature (see for example Pleskac, 2008; Rieskamp, 2008b; Rieskamp & Otto, 2006), we highlight the enhanced insight on risky behavior that mathematical models provide. Manuscript one tests which of several cognitive utility models best describes behavior in a lottery task. Manuscript two implements the BSR model to enhance the prediction power of the BART, and in manuscript three we challenge some of the assumptions of certain BSR variations.

Part I: Probing Risk Preference

Manuscript One: Losses in Short Behavioral Risk-Preference Measures

Schürmann, O., Andraszewicz, S., & Rieskamp, J. (2017). The Importance of Losses when Eliciting Risk Preferences. Manuscript submitted for publication.

As Bernoulli's EU theory suggests, people seem to under- or overweight values of decision outcomes. Different models can quantify this fact and estimate parameters specific to one individuals' utility of the actual values(Von Neumann & Morgenstern, 1945). Practitioners, as well as economic researchers alike, have an interest in quickly and accurately measuring risk-preferences in the form of individual utility function parameters. For this reason, a number of short measures exist, (supposedly) able to quantify individual risk parameters (Eckel & Grossman, 2002; Gneezy & Potters, 1997; Holt & Laury, 2002). In this manuscript, we first investigate a number of these measures and compare them to their counterparts: self-reported risk preference measures regarding retest-reliability and validity (Zuckerman, 2007; Zuckerman et al., 1978). Based on the several advantages of individual measures, we also introduce and compare a new behavioral risk preference measure based on lottery decisions, the BAR.

As mentioned in the Introduction, a major difference between many behavioral measures and self-report measures is the way risk is defined(Schonberg et al., 2011). While economic theories define risk as the variance of possible outcomes (Tversky & Kahneman, 1992; Von Neumann & Morgenstern, 1945), self-report measures often frame risk as the possibility of a loss. Outside of economic theories, practitioners and laypeople refer to risk more often as the possibility of a loss, and not as the variance of outcomes(March & Shapira, 1987). While these definitions are sometimes intertwined, many short lottery risk-measures do not incorporate losses at all, thus ignoring a crucial property of real-life risk-taking (e.g., Holt & Laury, 2002). A second difference is that self-report measures are typically framed as everyday life situations, while abstract lottery tasks are not. In the study presented in this manuscript, we introduce a newly developed lottery measure, the BAR, which is based on the assumptions of CPT and

thus includes loss gambles in addition to a real-life framing of the lottery-questions. We compare the performance of the BAR to a number of existing short lottery risk measures and self-report measures. Finally, we test a number of economic models of decision making to investigate the determinants of risky decision making using the BAR.

In an online study with 200 participants, we tested the same set of measures in two sessions, separated by one month. Testing the same measures with the same participants twice enabled us to provide some important test-criteria such as retest-reliability and validity of each measure. The set included three short behavioral lottery measures and three self-report measures of risk-taking (Barratt, 1965; Eckel & Grossman, 2002; Gneezy & Potters, 1997; Holt & Laury, 2002; Zuckerman et al., 1978). Importantly, we used a benchmark for real-life risk taking by including the general risk-question of the German socio-economic panel (SOEP-G). This question ("Are you generally willing to take risks?" has proven to be a very good predictor of real-life risk-taking in various studies (Dohmen et al., 2011; Frey et al., 2017).

The BAR comprises a set of six lottery questions, each framed as either a gambling decision in a casino or an investment decision on the stock market. Most importantly, the six different questions differ in the valence of the lotteries they present. While two question present gain-only lotteries, two present loss-only lotteries, and two questions include mixed lotteries (i.e., both gains and losses). In each question, the participant can decide between 14 options, ranging from a very safe to a very risky option. Each option has two outcomes occurring with a probability of .5 each. The BAR can be either used by averaging the raw choices of all six question or to estimate CPT parameters. While most short risk-elicitation measures are based on EU theory that does not distinguish between gains and losses, thus does not require the measurements of losses at all, the CPT accounts for differences between gains and losses (Köbberling & Wakker, 2005; Tversky & Kahneman, 1992).

To design the underlying structure of risk parameters of the BAR we assumed behavior in a lottery task to be best described by CPT. We tested this assumption by comparing a number of simple variations of the CPT model, including a model that was

akin to EU models–the expected shortfall model. The expected shortfall model takes into account only how much a decision maker falls short in their expectations about a possible outcome, to understand what factors drive behavior of participants in the BAR.

Our results show that based on the raw choices of each risk-measure, self-report measures are performing better than lottery tasks in both retest-reliability and correlation with real-life risk-taking as measured by correlation with the SOEP-G question. However, the BAR showed a significant improvement in retest-reliability compared to the existing short lottery measures. The BAR additionally also correlated better with the SOEP-G measure than the other lottery tasks, although this improvement was marginal.

Finally, by testing five models that incorporate different variations of cognitive, economic utility models, we provide strong support for the major influence of losses in decision making under risk. The CPT model with different utility parameters for gains, losses, and a loss-aversion parameter explained behavior in the BAR better than models that did not incorporate separate utility functions or loss aversion-parameters. Importantly, among the three CPT parameters, loss aversion showed the best retest reliability compared to the other two parameters. Finally, a single parameter of the expected shortfall model showed the best temporal stability as well as correlation with SOEP-G of all estimated risk-parameters.

Manuscript Two: Harnessing the Structure of Risk Preferences

Schürmann, O., Frey, R., Hertwig, R., Rieskamp, J., & Pedroni, A. (2017). Combining General and Specific Measures of Risk Preference Boosts Predictive Power in the Wild. Manuscript submitted for publication.

No everyday-situation in the "wild" has deadlier consequences than traffic decisions. More than one-fifth of people killed on roads worldwide are pedestrians (World Health Organization & World Health Organization, 2013). Pedestrian fatalities occur under various circumstances. Previous research has studied the extent to which individual risk-preferences determine the risks taken by pedestrians and other actors in

traffic. However, these studies commonly probed risk-preferences and behavior concurrently and measured risky traffic behavior in a laboratory setting rather than in the wild. In this manuscript, we report the findings of a study that predicted risky real-life street crossing decisions by harnessing the complementary predictions of domain-specific as well as general measures of risk-preferences.

A crucial element in predicting actual risk-taking behavior is how to best conceptualize and measure the construct of risk-preference (Appelt et al., 2011; Fox & Tannenbaum, 2011; Frey et al., 2017; Schonberg et al., 2011). Some researchers argue that risk-preference is a one-dimensional construct that manifests consistently in behavior across different domains and circumstances (Jackson, Hourany, & Vidmar, 1972). Another view is that risk-preferences vary across situations and domains (Weber et al., 2002). Finally, a third and more recent view is that risk-preferences–similar to intelligence (Spearman, 1904)–have both general and specific components (Frey et al., 2017; Highhouse et al., 2016). We harness these recent findings of the structure of risk-preference and test a set of measures, both specific and unspecific. We test the predictive abilities of these measures of risky, naturalistic street-crossing behavior after a delay of ten months. We additionally employed mathematical cognitive models of risk-taking to extract the best prediction of a behavioral risk-preference measure (i.e., the BART). We tested the predictions of the different risk preference measures on two criteria; both conducted in a second session which, on average, took place over ten months (at least six months) after the preference testing. The two criteria were (a) averaged decisions in a naturalistic street crossing task and (b) self-reported dangerous driving behavior (Reason, Manstead, Stradling, Baxter, & Campbell, 1990). For risky decision making in street crossing, we developed a novel naturalistic task. In the task, participants were standing on the side of a street and making street-crossing decisions in front of a number of arriving cars or trams.

The results showed that the domain-general risk-preference measures were not correlated with the domain-specific measure, but all measures were substantially correlated with the target criterion of risky street crossing decisions. Consequently, the

best composite models predicted behavior substantially better than any single predictor model. This finding is in line with recent findings in other fields, where combining independent predictors increased predictive accuracy substantially (Armstrong, 2001; Menkhoff & Sakha, 2017). The best model was the one combining a traffic-specific component and a general personality measure (sensation seeking). It predicted a substantial amount of the variance in street-crossing behavior ($R^2 = .35$) and was also able to predict self-reported dangerous driving behavior, thus showing evidence of generalizability. Our results support the tendency that risk-preference has both domain-general and domain-specific components (Frey et al., 2017) and both aspects in combination determine risk-taking behavior in an actual street setting after an average of ten months.

Part II: Probing Risk Perceptions

Manuscript Three: Risk perceptions: Their impact on risk taking and the effect of early experience

Schürmann, O., Pleskac, T. J., Frey, R., (2017). Risk perceptions: Their impact on risk taking and the effect of early experience. Manuscript submitted for publication.

In the second part, we loosen the focus on (stable?) risk preferences and instead turn to risky decision making as determined by cues of the environment. As mentioned above, many decisions in everyday life are not based on precise descriptions of choice properties, but are perceived (willingly or unwillingly) through experiences (Hertwig & Erev, 2009). The question that arises is to what extent the perception of choice properties determines people's engagement in risky decision making?

In the literature about self-report risk preference measures, people have been shown to engage in risky decision making in line with how they rate the riskiness (i.e., chance of a loss or harm) of a behavior(Weber & Hsee, 1998). Additionally, people making decisions from experiences are believed to learn about probabilities of outcomes (Hertwig & Erev, 2009; Rieskamp, 2008a). In hopes of understanding the influence of early experiences on risk perceptions and how they, in turn, determine behavior, we

gauge risk perceptions in a sequential risk-taking task, the BART (Lejuez et al., 2002).

During the BART, participants accumulate points for pumping up several virtual balloons (Lejuez et al., 2002). Their task on each trial is to pump up the trial's balloon and decide when to stop pumping. If they stop before the balloon explodes they can keep the points, but if the balloon explodes before they decide to stop, they lose the points. The average number of pumps on non-exploding balloons (adjusted BART score) is then used as a risk-taking measure and has been found to correlate with real-world risky behaviors like drug use and engagement in unprotected sex (see Lauriola et al., 2013, for a review). The reason why the BART is particularly interesting to study dynamic risk-perceptions is its stochastic nature that is well-defined. Participants are purposely kept uninformed about the probabilities in the BART (Pleskac, 2008). In two experiments, we tested how risk-perceptions (in the form of probability ratings) correlate with pumping behavior in the task and how these perceptions were influenced by initial experiences in the task.

In Study 1, after completing the BART, participants indicated their risk-perception by rating the probability of an explosion *on the next pump* for various sizes of balloons. We found that participants' perceptions of this probability neither corresponded to the actual structure of the task nor the assumptions of cognitive models of choice behavior. Instead, participants estimation of the probability increased in a sigmoid fashion with pumping and reached certainty much earlier than the true structure of the task (where the balloon explodes with on the 128th pump). Moreover, the results showed that the changing point of the modeled probability-rating curve (i.e., the point where the sigmoid function inferred from the individual probability ratings reached an explosion probability of .5) correlated strongly with the average number of pumps made.

In Study 2, we investigated how the initial experience of an explosion influenced risk perception and risky decision making in the task. For this, we varied the explosion point of the first balloon for three between-subject groups (after 64, 16, and 96 pumps for the control group, low group, and high group, respectively). Additionally, we probed

the probability ratings twice, once immediately after the first trial, and once after the task. With this, we aimed to investigate the effect of the manipulation on the probability rating and to examine the change in perception over the task.

Our results strikingly showed how risk perceptions are dependent on the initial experience in the task. Participants differed in the first ratings of perception (all three groups indicated significantly different probability ratings in the first rating). Moreover, the different initial experiences altered the behavior throughout the task, such that the adjusted BART score was significantly lower in the low group. The high group also showed a trend to be higher than the control group. A mediation analysis confirmed that the effect of the first explosion experience onto the average BART score was significantly mediated by the perception of the explosion probability. These results corroborate the important role that risk-perception plays in the BART and the plasticity of these perceptions to early experiences within the task. They further challenge existant approaches that infer risk perceptions (i.e., subjective probabilities) from observed choice behavior using cognitive models. These models were not able to correctly incorporate the influence of the early experience onto subjective probabilities. Finally, the results speak to the recent puzzles of measuring risk-preferences using behavioral laboratory task.

Summary

What are the determinants of risky decision making in the wild? Moreover, how, if at all, can we coherently measure and investigate the determinants of risky decision making in everyday-life situations? The findings of the three presented manuscripts harness the potential of probing risk-taking behavior with various laboratory measures in combination with mathematical modeling of cognitive processes. While part one walks along the path of Daniel Bernoulli and uncovers how risk-preferences can be probed to better uncover relevant determinants in naturalistic decisions, part two investigates how decisions under uncertainty cause people to base their risk-preferences on the perceptions of cues in the environment.

Part one shows, when measuring risk preferences to determine real-life risk taking, special care needs to be taken how preferences are measured. Manuscript one shows that the presentation of losses and the framing of risk in decisions not only improves measurement test criteria but helps to bridge the gap between different definitions of risks. Manuscript two shows that, in order to maximize the predictability of real-life risk taking through risk-preferences, it is crucial to consider the structure of risk preferences, such as those specific to situational factors in addition to those that are more general and unspecific to a certain situation.

Part two of the dissertation sheds some light onto the gap between decisions in which probabilities are unknown, by outlining the determining influence of perceptions onto behavior, as well as the constructions of these perceptions by initial experiences.

The current work thus improves our understanding of how risk-preferences can be measured, but also how the cognitive mechanisms underneath might influence behavior in real-world situations. With this, we bridge the gap between the lab and the wild and provide crucial additions to the field of theoretical- as well as applied decision research. Thereby, and especially thanks to its everyday-life connection, the findings presented in this dissertation help to educate individuals in making better decisions.

References

Appelt, K. C., Milch, K. F., Handgraaf, M. J., & Weber, E. U. (2011). The Decision making individual differences inventory and guidelines for the study of individual differences in judgment and decision-making research. *Judgment and Decision Making*, *6*(3), 252.

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners* (Vol. 30). Springer Science & Business Media.

Barratt, E. (1965). Factor analysis of some psychometric measures of impulsiveness and anxiety. *Psychological Reports*, *16*(2), 547 –554.

Bernoulli, D. (1738/1954, January). Exposition of a New Theory on the Measurement of Risk. *Econometrica*, *22*(1), 23. doi: 10.2307/1909829

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432–459. doi: 10.1037/0033-295X.100.3.432

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

Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, *23*(4), 281–295. doi: 10.1016/S1090-5138(02)00097-1

Fox, C. R., & Tannenbaum, D. (2011). The Elusive Search for Stable Risk Preferences. *Frontiers in Psychology*, *2*. doi: 10.3389/fpsyg.2011.00298

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk Preference shares the structure of major psychological traits. *Under Review*.

Gigerenzer, G. (2006). Out of the Frying Pan into the Fire: Behavioral Reactions to Terrorist Attacks. *Risk Analysis*, *26*(2), 347–351. doi: 10.1111/j.1539-6924.2006.00753.x

Gneezy, U., & Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, *112*(2), 631–645.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (n.d.). Decisions from experience and the effect of rare events in risky choice. , *15*(8), 534–539.

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523. doi: 10.1016/j.tics.2009.09.004

Highhouse, S., Nye, C. D., Zhang, D. C., & Rada, T. B. (2016, January). Structure of the Dospert: Is There Evidence for a General Risk Factor? *Journal of Behavioral Decision Making*, n/a–n/a. doi: 10.1002/bdm.1953

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

Jackson, D. N., Hourany, L., & Vidmar, N. J. (1972). A four-dimensional interpretation of risk taking1. *Journal of personality*, *40*(3), 483–501.

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

Knight, F. H. (1921). Risk, uncertainty and profit. *Houghton Mifflin Company*.

Köbberling, V., & Wakker, P. P. (2005). An index of loss aversion. *Journal of Economic Theory*, *122*(1), 119–131. doi: 10.1016/j.jet.2004.03.009

Lauriola, M., Panno, A., Levin, I. P., & Lejuez, C. W. (2013, January). Individual Differences in Risky Decision Making: A Meta-analysis of Sensation Seeking and Impulsivity with the Balloon Analogue Risk Task: Personality and Risky Decision Making. *Journal of Behavioral Decision Making*, *27*(1), 20–36. doi: 10.1002/bdm.1784

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., ... Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84. doi: 10.1037//1076-898X.8.2.75

March, J. G., & Shapira, Z. (1987). Managerial Perspectives on Risk and Risk Taking. *Management Science*, *33*(11), 1404–1418.

Menkhoff, L., & Sakha, S. (2017). Estimating risky behavior with multiple-item risk measures. *Journal of Economic Psychology*, *59*, 59–86. doi:

10.1016/j.joep.2017.02.005

Ore, O. (1960). Pascal and the Invention of Probability Theory. *The American Mathematical Monthly*, *67*(5), 409–419. doi: 10.2307/2309286

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

Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 167–185. doi: 10.1037/0278-7393.34.1.167

Ratcliff, R., & Smith, P. (2015, April). Modeling Simple Decisions and Applications Using a Diffusion Model.
doi: 10.1093/oxfordhb/9780199957996.013.3

Reason, J., Manstead, A., Stradling, S., Baxter, J., & Campbell, K. (1990). Errors and violations on the roads: a real distinction? *Ergonomics*, *33*(10-11), 1315–1332.

Rieskamp, J. (2008a). The importance of learning when making inferences. *Judgment and Decision Making*, *3*(3), 261.

Rieskamp, J. (2008b). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1446–1465.

Rieskamp, J., & Otto, P. E. (2006). SSL: A Theory of How People Learn to Select Strategies. *Journal of Experimental Psychology: General*, *135*(2), 207–236. doi: 10.1037/0096-3445.135.2.207

Schonberg, T., Fox, C. R., & Poldrack, R. A. (2011). Mind the gap: bridging economic and naturalistic risk-taking with cognitive neuroscience. *Trends in Cognitive Sciences*, *15*(1), 11–19. doi: 10.1016/j.tics.2010.10.002

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. doi: 10.2307/1884852

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292. doi: 10.2307/1412107

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*(4), 297–323.

Von Neumann, J., & Morgenstern, O. (1945). *Theory of games and economic behavior*. Princeton University Press Princeton, NJ.

Von Neumann, J., & Morgenstern, O. (2007). *Theory of games and economic behavior*. Princeton, UK: Princeton university press.

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling Behavior in a Clinically Diagnostic Sequential Risk-Taking Task. *Psychological Review*, *112*(4), 862–880. doi: 10.1037/0033-295X.112.4.862

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*(4), 263–290. doi: 10.1002/bdm.414

Weber, E. U., & Hsee, C. (1998). Cross-Cultural Differences in Risk Perception, but Cross-Cultural Similarities in Attitudes Towards Perceived Risk. *Management Science*, *44*(9), 1205–1217. doi: 10.1287/mnsc.44.9.1205

World Health Organization, & World Health Organization (Eds.). (2013). *Global status report on road safety 2013: supporting a decade of action*. Geneva, Switzerland: World Health Organization.

Zuckerman, M. (2007). Sensation Seeking and Risk. In *Sensation seeking and risky behavior*. (pp. 51–72). Washington: American Psychological Association.

Zuckerman, M., Eysenck, S. B., & Eysenck, H. J. (1978). Sensation seeking in England and America: cross-cultural, age, and sex comparisons. *Journal of consulting and clinical psychology*, *46*(1), 139.

The Importance of Losses when Eliciting Risk Preferences

By Oliver Schürmann, Sandra Andraszewicz, Jörg Rieskamp *

Researchers and practitioners have developed a number of methods to quickly assess people's individual risk preferences in financial decision-making. Commonly, two types of elicitation methods can be distinguished; behavioral risk-preference elicitation methods and self-report methods. Behavioral methods, contrary to self-report methods, offer the possibility to estimate risk preference parameters of economic theories. On the other hand behavioral measures show relatively low retest reliability and ecological validity, in comparison to the self-report measures. In this paper, we examine the retest reliability of various risk-preference elicitation methods and accesses their external validity. Next, propose a new measure – BAR, which is behavioral-based with questions framed in a real-life context that mimic self-report type of measures. BAR goes beyond previous measures by distinguishing risk aversion and loss aversion. By rigorously testing BAR against selected existing risk measures, we show that incorporating loss aversion in a risk measure increases its reliability. In sum, this study illustrates major challenges when eliciting people's risk preferences and proposed new directions in eliciting people's propensity to take risks in finance.

JEL: D81; D91; G4; G11

Keywords: "risk profiling; risky decision-making; utility curve; loss aversion"

* Schürmann: University of Basel, Missionsstrasse 62a, 4055 Basel, o.schuermann@unibas.ch. Andraszewicz: Swiss Federal Institute of Technology (ETHZ), Clausiusstrasse 59, 8092 Zürich, sandraszewicz@ethz.ch. Rieskamp: University of Basel, Missionsstrasse 62a, 4055 Basel, joerg.rieskamp@unibas.ch. We would like to thank the master project in economic psychology 2016 for developing earlier versions of the BAR.

Following the great recession during the late 2000s, financial regulatory agencies worldwide introduced directives protecting investors in financial markets. Accordingly, financial institutions nowadays are required to assess people's risk preference or risk tolerance before offering investment products. For example, according to the rule of suitability included in the Article 19(4) of Markets in Financial Instruments Directive (MiFID) issued by the European Commission,[1] a financial institution offering investment advice and products is obliged to collect necessary information about the risk profile of a customer. Similar regulation holds for US-based companies, as defined by the Financial Industry Regulatory Authority (FINRA)[2]. Despite minor differences, most regulations require firms to assess the risk preference of their client (Boskovic et al., 2010), which resulted in extensive efforts among practitioners to measure people's risk preferences appropriately.

Whereas practitioners face legal requirements regarding measuring people's risk preferences, researchers are particularly interested in risk preferences as the major component of economic theory of decision making under risk (Charness et al., 2012; Crosetto and Filippin, 2016; Pedroni et al., ress). Therefore it is not surprising that a plethora of various risk-preference elicitation methods (EM) for identifying people's risk preferences exist. However, it is less clear which risk-preference EM offers the most reliable results (Pedroni et al., ress).

There are several important criteria for assessing people's risk preferences. First, it is important that the method gives similar results at different points in time, reflecting the reliability of the estimate. It appears to be a reasonable assumption that people's risk preferences should be stable over a relatively short time period such as weeks or months so that elicited risk preferences at two time points should be highly correlated. However, people's risk preferences might change over longer time periods as risk preferences change over the lifespan (Josef

[1] http://ec.europa.eu/internal_market/securities/docs/isd/dir-2004-39-implement/dir-6-2-06-final_en.pdf

[2] http://finra.complinet.com/en/display/display_viewall.html?rbid=2403&element_id=3638&record_id=4315&filtered_tag=

et al., 2016). If a method leads to low reliability, it may be caused by unsystematic measurement error due to people's lack of understanding of the task used for measuring their risk preference (Dave et al., 2010).

Second, a good risk-preference EM should not only hold for the experimental laboratory but should also correctly predict people's "behavior in the wild". Schonberg et al. (2011) identified the divergence between real-life risk-preference EMs and more theoretically motivated methods using relatively abstract decision contexts, such as monetary gambles. Their results show that different neural correlates are at play during static risk-preference EM tasks than during emotionally affective tasks. In particular, Schonberg et al. (2011) point out the common gap in risk definitions. Whereas behavioral psychologists and economists define risk as the variance of outcomes, laypeople and practitioners tend to refer to risks as the possibility of losses (March and Shapira, 1987). This gap between the naturalistic and economic theoretical EMs directly maps onto the external validity difference between the more realistic and more abstract tasks.

Third, from a practical perspective, it is desirable to measure people's risk preferences with a quick and easy tool. Some EMs take up to 15 minutes to be completed (i.e., Lejuez et al., 2003; Weber et al., 2002) whereas other measures require up to only 2 minutes (i.e., Holt et al., 2002; Eckel and Grossman, 2008a). Practitioners strongly lean towards fast methods due to various time constraints in business applications.[3] Also, less complex EMs are easier for participants to understand and result in less noisy measurements (Dave et al., 2010).

Finally, Charness et al. (2012) indicate that being able to estimate risk preference parameters that correspond to theoretical decision-making principles helps using and comparing the measure within a theoretical economic framework. Schonberg et al. (2011) list the possibility of decomposing one's risk preference to economic theoretical parameters as a requirement for a good risk-preference EM.

[3]In the industry, a commonly accepted upper time limit for eliciting a customer's risk preference is 5 minutes.

The ability the quantification of one's risk preference enables the objective classification of risk-averse, risk-neutral and risk-seeking people, as well as directly integrating one's risk profile in financial investment models. Therefore, it is desirable to be able to estimate one's individual *risk parameter*, which defines the curvature of one's utility function.

We use these four criteria to set the framework for evaluating risk-preference EMs. The goal of this paper is threefold. First, we compare the different risk-preference EMs with each other, illustrating their pros and cons. We evaluate the test-retest reliability of representatives of existing risk-preference elicitation methods for risky financial decision-making over a period of one month. Second, we assess their external validity by examining how well these measures correlate with self-reported real-life risk-taking. Third, we propose a new risk preference-elicitation method, the Basel Risk Preference (BAR) measure. The BAR has the advantage of distinguishing between degrees of risk preferences and loss aversion, by asking people questions about gains and losses. By rigorously testing various quantitative models of decision making following the concepts of cumulative prospect theory (CPT) (Tversky and Kahneman, 1992) against each other, we show the importance of differentiating between risk preferences and loss aversion.

The paper is structured such that in Section I, we describe different types of risk-preference elicitation methods. We emphasize advantages and disadvantages of each method. Section II presents our selection of the representative risk-preference EMs that we test for test-retest reliability and correlation with real-life risk-taking. In Section III, we outline the newly proposed measure and the choice models that we tested for evaluating the importance of loss aversion in measuring risk propensity. Section IV outlines the experimental procedures, Section V presents the empirical findings, while Section VI concludes.

I. The Pros and Cons of Different Types of Risk-Preference Elicitation Methods

A. Lottery-based risk-preference elicitation methods

A large number of different risk-preference elicitation methods can be classified into behavioral and self-report EMs (Frey et al., ress). In many of the behavioral EMs, participants choose between lotteries. A popular lottery-based way of measuring risk preferences is to ask people to make repeated choices between a number of pairs of lotteries (Wakker, 2010; Murphy and ten Brincke, 2017). Each lottery is characterized by different outcomes and the corresponding probabilities with which each outcome occurs. Lottery-based EMs have the advantage that the choice options can be precisely described and quantified, for instance, by the expected value and variance of the lotteries' outcomes.

When the goal is to measure people's risk preferences quickly, practitioners might prefer more concise methods over more precise but time- and effort-intense measures (Abdellaoui et al., 2011). Various risk-preference EMs exist that require people to make a limited number of choices between risky options. For instance, one of the most prominent risk-preference EMs proposed by Holt et al. (2002, HL) requires people to make only 10 choices between a pair of risky gambles. Participants successively move down a list of pairs of two-outcome lotteries, always choosing one of the options they prefer. The pairs are constructed such that one lottery has a lower variability of outcomes in comparison to the other lottery. Over the 10 pairs, the outcome probabilities are changed such that a risk-neutral decision maker will prefer the low-variance lottery for the first six pairs and then switch to the higher-variance lottery due to its higher expected value. Thus, whether this "switching point" occurs before or after the sixth pair of lotteries is a measure of one's risk preference.

Despite its shortness, the structure and the variation of probabilities in the HL task has often been shown to be cognitively demanding (Dave et al., 2010).

Therefore the HL risk-preference measure can be biased by the cognitive ability of the participants. Also, due to the fact that people often process probabilities in a non-linear fashion (Abdellaoui et al., 2011; Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), the risk-preference measurement with the HL method might be biased by the perception of probabilities.

A simplified method by Eckel and Grossman (2008b, cf., EG task) requires the choice of one out of six two-outcome lotteries. In contrast to Holt et al. (2002), the variation in the options is obtained through manipulation of the outcomes of each lottery, keeping the probability of each outcome fixed at 50%. This way, probability weighting would not influence the expected utility of the gambles. Dave et al. (2010) argue that the EG task is easier to understand than the HL task and leads to fewer inconsistent choices.[4] On the other hand, the EG task measures the risk preference parameter of an expected utility theory with less precision than the HL task (i.e. a larger range of the estimated risk parameter value) due to the smaller number of choices that a participant has to make. Additionally, the EG task only allows measuring risk-averse risk preferences whereas risk-seeking preferences cannot be distinguished because only one choice option corresponding to risk-seeking is provided.

Another even simpler lottery-based risk-preference EM by Gneezy and Potters (1997, cf., GP task) requires people to make only a single investment decision. Participants have to decide how to allocate a given amount between cash and a risky investment that yields 2.5 times the amount invested with a probability of 50% or a total loss of the invested amount with the remaining 50% probability. The GP measure poses very little demand on numerical abilities and is thus easy to understand. However, this measure has the disadvantage that it cannot capture risk-seeking preferences because a risk-neutral decision maker would invest the whole endowment into the investment option (Charness et al., 2012).

In general, the different lottery-based EMs (including the HL, GP, and EG

[4]In the HL task, inconsistent choices refer to multiple switching points between options A and B.

tasks) have the advantage that the choice options are precisely quantified, so people's choices can be used to estimate the parameters of an expected utility theory that describes people's risk preferences. Depending on the specific expected utility theory model, one or more parameters represent people's risk preferences. Furthermore, due to the clear quantification, lottery-based EMs allow for incentivizing participants' choices by paying all or a part of the outcomes of the decisions to the participant. In economic thinking, providing incentives for choices and outcomes may lead to more meaningful and reliable choice behavior (Charness et al., 2016).

However, lottery-based risk-preference EMs have recently been criticized for their relatively low retest reliability and low external validity compared to self-report measures (Frey et al., ress; Pedroni et al., ress). Specifically, Lönnqvist et al. (2015) found that low retest correlation for the Holt and Laury task ($\rho = .56$,$p > .1$) but a large correlation for the SOEP-G measure ($\rho = .76$,$p < .001$) after one year. Other studies find that lottery-based EMs show some temporal stability after a year (HL, EG Galizzi et al., 2016) and after six months (HL, others Frey et al., ress). Self-report measures in general (i.e., SOEP-G, SS, IMP) usually provide much stronger stability over time.

Three challenges that arise with most lottery-based risk-preference EMs might explain the reasons for this criticism. First, the precision of the estimated risk parameter can be low when using only one or a few decisions to estimate the parameter(s) of a utility function (Crosetto and Filippin, 2016). Depending on the structure and length of the task, the range of the estimated parameter(s) might be relatively large (Crosetto and Filippin, 2016). The lower the number of choices available, the larger the measurement error of the parameter estimate. For example, the HL task can classify people in 10 different risk preference categories, while the greater simplicity of the EG task comes at the price of a coarser estimation allowing only six categories.

Second, the scarcity of choices used to estimate risk parameters may result in

high measurement error due to the probabilistic nature of preferential choices (Rieskamp, 2008). Despite having a preference of option A over option B, it can happen that a decision maker does not choose option A in 100% of the cases, but in, for example, 75% of the cases. Thus, basing the risk-preference estimate on a single choice increases the chance of measurement error in comparison to aggregating multiple choices. Indeed, recent research indicates that aggregating multiple risk-preference estimates can boost reliability and predictive power of people's risk-taking behavior outside the laboratory (Menkhoff and Sakha, 2014; Frey et al., ress). However, what is unclear in the lottery-based tasks is whether risk taking in these lotteries is stimulating (i.e., is treated as an entertaining game) or is instrumental (i.e., has a long-term investment goal, Zaleśkiewicz, 2001). Not distinguishing between the goal of risk-taking (stimulating vs. instrumental) may result in inaccurate or erroneous risk-preference estimation for investment situations.

The third challenge of the proposed EMs is that they often do not examine choice options including losses. Treating losses and gains differently is one of the major violations of the standard economic view of expected utility maximization (Köbberling and Wakker, 2005). Tversky and Kahneman (1992) report that people overweight losses relative to gains, such that the utility curve for losses is steeper than it is for gains. In simple terms, losing 10$ has a bigger impact than the opposite of gaining 10$. This *loss aversion* phenomenon was reported in many empirical studies (see Köbberling and Wakker, 2005; Kahneman and Tversky, 2000, for example). Loss aversion can also explain a variety of decision patterns observed in real-life situations (see Camerer, 1998, for an overview).

Also, laypeople and practitioners more often think about risk regarding possible losses or negative events, rather than the variance, as it is assumed in theoretical economic or finance theories (see Schonberg et al., 2011, for an elaborate discussion on the definitions of risk). In this line, even experienced managers tend to see risk as possible losses, rather than the variance of outcomes (March and

Shapira, 1987). Thus, if people perceive risk as the potential of losses but a task only varies the variance of positive outcomes, then people can hardly express their risk preferences. Furthermore, according to prospect theory (Tversky and Kahneman, 1992), people tend to be risk averse in the gain domain but risk seeking in the loss domain. Thus, people's risk preferences are accordingly numeric-domain specific. This implies that EMs following the lottery-based approach, which only uses lotteries in the gain domain, may not accurately reflect people's perception of risk. This could also be a reason why these measures might not perform well in predicting people's risk-taking behavior in many real-life situations (Schonberg et al., 2011).

B. Self-report based risk-preference elicitation methods

Self-report EMs are a different type of measure that is commonly used by psychologists and psychometricians. These measures have recently been shown to have better retest reliability, as well as external validity than lottery-based EMs (Frey et al., ress).

Most of the self-report measures use various multi-item questionnaires asking people about hypothetical decision situations where a person has to decide on how s/he would behave in that situation. For example, in the Sensation Seeking Scale introduced by Zuckerman (2007), people repeatedly have to evaluate which of two statements characterizes their own opinion most accurately. One of these statements reprdesents a high risk-seeking tendency ("I am thinking about going bungee jumping one day") or a risk-averse tendency ("I prefer the comfort of my own home"). People's risk- or sensation-seeking preferences are then characterized by the proportion of statements where the high-risk/sensation-seeking tendency was selected. The sensation-seeking score has been shown to correlate with numerous risky behaviors, including risky gambling behavior (Wong and Carducci, 1991) or risky alcohol use (Hittner and Swickert, 2006).

Similarly, the Barratt Impulsiveness Scale asks participants about how likely

they would be to engage in an impulsive or less impulsive behavior (Barratt, 1965). The personality trait of impulsiveness has been shown to be connected to risky behavior in many areas such as gambling (Mishra et al., 2010) and to generally influence reward sensitivity (Penolazzi et al., 2012), which makes it a crucial factor and predictor of risky decision-making in any situation.

Another prominent example of a self-report risk-preference measure is used in the German Socio-Economic Panel (SOEP) Study (Wagner et al., 2007). In the SOEP, participants are asked a number of questions about how much risk they take or are willing to take in different real-life contexts. For instance, participants are asked to what extent they are willing to invest their money in a risky stock. In that sense, SOEP is a questionnaire that does not require participants to answer questions about hypothetical situations but instead, it asks about their real-life experiences. Indeed, Dohmen et al. (2011) found that a question asking participants about their general willingness to take risk (SOEP-G) was the best overall predictor for real-life risk-taking behavior such as holding stocks, risky occupational choices, or substance abuse. Therefore, SOEP-G can be reliably used as an indication of one's real-life risk-taking behavior.

Despite their advantages, self-report EMs have been criticized for providing no incentives to the participants because all answers are hypothetical and do not involve any real consequences (see Dave et al., 2010, for a more elaborate discussion). Therefore, self-report EMs are not incentive compatible, making it questionable whether they yield honest and economically motivated answers. Self-reports could be considered as "cheap talk" because participants could report an idealized view of themselves or answer questions according to social desirability. Given this reasonable criticism, it is astonishing that the retest reliability of self-report measures is relatively high, ranging between .7 and .8 (Frey et al., ress). One explanation could be that this high reliability results from a stable image of self that a respondent has.

Another major disadvantage of the self-report EMs is that they only provide

ordinal measures of people's risk preferences. Thereby self-report EMs do not allow to quantify people's risk preferences relative to a specific expected utility theory of decision-making and its specific risk preference parameters. The ordinal measure of people's risk preferences with self-report measures often do not even allow the quantification of whether a person is risk-averse or risk-seeking which, of course, is an important theoretical and practical distinction.

II. Selection of Measures for Reliability and External Validity Evaluation

In this study, we evaluate the test-retest reliability and external validity of selected risk-preference elicitation methods. We include three short lottery tasks: the Holt and Laury task (cf. HL, Holt et al., 2002) as the most commonly used short multiple price list measure, the Eckel and Grossmann task (cf. EG, Eckel and Grossman, 2008b) as a short measure that keeps outcome probabilities constant at the value of 50%, and the Gneezy and Potters task (cf. GP, Gneezy and Potters, 1997), which is a one-question investment task that makes it simple and easy to understand. All these measures are 1) appropriate for measuring risk propensity in the finance domain, 2) can be incentivized, 3) are short, and 4) allow for estimating a person's risk preference parameter in an expected utility model. Additionally, we include the Sensation Seeking Scale (cf. SS, Zuckerman, 2007) and Barratt Impulsiveness Questionnaire (cf. IMP, Barratt, 1965), which have been shown to have high inter-measurement correlations and to correlate highly with each other and other risk EMs (Frey et al., ress).[5].

Each of these methods has some desirable features. By having 10 choice options, HL enables a more precise estimation of the *risk parameter* than other measures. EG keeps the probabilities of outcomes constant at the level that should be the easiest to capture for the respondents (i.e., 50%). Finally, in GP, the value of the endowment can be freely chosen by the researcher, which makes the measure more

[5]We did not include another prominent self-report measure, the Domain Specific Risk Attitude Scale (DOSPERT), because it was shown to be highly correlated with the sensation-seeking scale (Frey et al., ress)

general and applicable for different purposes. However, it is less clear how reliable the measures are and how good they are in predicting people's real-life risk-taking so we will examine this question in the following empirical study. Furthermore, we suggest a new method of eliciting people's risk preferences by combining elements from the behavioral lottery-based measurement approach with elements from the self-report measurement approach. The new measure also includes the loss domain for distinguishing risk preferences from loss-aversion preferences.

III. Basel Risk Preference (BAR) Measure

On the basis of the advantages and disadvantages of the lottery-based and the self-report EMs, we suggest a new risk-preference measure that represents a combination of both approaches. First, this method requires people to choose repeatedly between clearly defined risky choice options. Each option can be represented as a lottery with two outcomes that occur with a probability of 50%. To allow precise quantification of people's risk preferences, the EM has six items, each with fourteen choice options. Furthermore, two items entail gains, two entail losses, and two entail loss and gain outcomes (i.e., mixed lotteries). By including losses and gains, it is possible to identify not only people's risk preferences but also potential loss-aversion preferences. Second, to make the choices less abstract, the options are not presented as lotteries but are embedded in a real-life choice context similar to that found in self-report EMs. By providing a real-life context, people should have fewer difficulties in understanding the choice situation, implying fewer cognitive demands for performing the task accurately.

More precisely, the BAR uses either an investment situation or a casino gambling situation as a real-life context. In the investment situation, people are told that they invested money in a stock market and they are provided with their current portfolio value. They are asked which of different investment options they prefer. In the casino setting, people are put in a hypothetical situation in a casino with a certain balance value. They are presented with different lotteries

and can choose one of them. Thus, the casino situation explicitly provides people with choices among lotteries that are embedded in a real-life context of a casino, possibly fostering understanding of the question.

In the BAR, three questions with larger values (i.e., about +/-1000 – 3700 units of the experimental currency) use the investment situation, and three other questions with smaller values (i.e., +/- 10 – 300 units of experimental currency) use the casino situation. The values of the outcomes were chosen such that they match the nominal money values in equivalent real-life situations. Also, using two different range values enables for the fact that people's risk preferences may differ depending on the values of stakes (i.e., small vs. large) to be taken into account.

The outcomes of the specific lotteries of all six items have been specified such that they allow the estimation of utility function parameters of cumulative prospect theory (cf. CPT Tversky and Kahneman, 1992; Kahneman and Tversky, 1979). In CPT, the utility $u(x)$ of an outcome x distinguishes between losses and gains:

$$u(x) = \begin{cases} x^{\alpha} & , \quad \text{for} \quad 0 \leq x \\ -\lambda \times -x^{\beta} & , \quad \text{for} \quad x < 0 \end{cases}, \tag{1}$$

where α is the utility-function parameter for gains, β is the parameter for losses, and λ represents a loss-aversion parameter. The loss-aversion parameter defines how much an individual over-weights losses compared to gains. The two items of the BAR in the gain domain are used to estimate α, the two items in the loss domain are used to estimate β, and the two items with mixed outcomes are used to estimate λ.

The 14 choice options for each of the gain items are structured such that a single option should be chosen by a risk-neutral decision maker as it offers the highest expected value of all 14 options. Additional eight options have a lower

expected value than the risk-neutral option, but offer the advantage of lower risk defined as the variance of the outcomes. Therefore, these eight options allow the identification of people's risk aversion. The eight options correspond to α-values ranging between 0 and 1 (see Table A.A1 in Appendix A for the detailed information about the α-values corresponding to each choice option). The additional five choice options also have a lower expected value than the risk-neutral option but offer the opportunity of high outcomes due to higher risk; that is, the higher variance of outcomes. Therefore, these five choice options, allowing the identification of people's risk-seeking preferences. Specifically, the risk-seeking options correspond to α-values ranging between 1 and 2. The number of risk-averse and risk-seeking choice options is asymmetric because we assumed that most people are risk-averse (Pratt and Zeckhauser, 1987).

The structure of the loss items is similar to that for the gain items. There is one option with the highest expected value that should be chosen by a risk-neutral decision maker. However, following the assumptions of CPT we assume that in the loss domain, the majority of the people have risk-seeking preferences (Tversky and Kahneman, 1992). Therefore, there are an additional eight options, for which the expected value decreases and the risk (defined as the variance of the outcomes) increases, allowing the identification of risk-seeking preferences. The utility parameter β corresponding to these choice options ranges from 0 to 1. The remaining five choice options have lower expected value than the risk-neutral option and decreasing risk (i.e., variance), allowing the identification of risk-averse preferences. The corresponding parameter β ranges between values of 1 and 2 (see Table A.A1 in Appendix A).

Finally, two items had lotteries with positive and negative outcomes (i.e., the mixed domain). These mixed lotteries allow the estimation of people's loss-aversion preferences. However, for estimating the loss-aversion parameter, it is necessary to specify the risk preferences parameters α and β. Therefore, the 14 pairs of outcomes of the mixed lotteries were constructed by assuming a value

of 0.90 for both α and β, which are the median values assumed in the literature (Köbberling and Wakker, 2005). When people's risk preference deviates from these assumed values, the mixed items can still be used to estimate people's loss-aversion preferences, but the estimates will become less accurate. To avoid this inaccuracy, a utility model can also be estimated, as illustrated in Section III.A.

For the two items in the mixed domain, one option should be chosen by a person that is not loss averse. An additional 10 choice options were constructed so that they should be chosen with increasing loss-aversion preferences, corresponding to loss-aversion parameter values between 1 and 3.2. The remaining two choice options allow identification of the opposite of loss aversion, namely gain-seeking preferences corresponding to loss-aversion parameters between 0.5 and 1 (see Table A4 for detailed information about the relation between the lotteries and their corresponding λ-parameters).

The BAR is constructed such that each choice allows identification of one of the three risk-preference parameters of the cumulative prospect theory. However, each choice does not allow a point-estimate of the parameter value, but instead it is consistent with a specific range of parameter values. This range of parameter values is 0.1 for each choice for α and β and about 0.2 for λ. This method of identifying the parameters of the utility function is similar to the method suggested in Holt et al. (2002). However, because people are asked to make two choices each in the gain, loss, and mixed domain, the pair of choices in each domain might not be fully consistent. Therefore, pragmatically, the average of the corresponding parameter values could be selected. The loss-aversion parameter λ represented in Table A4 in Appendix A only holds for the assumed α and β parameters and might, therefore, be error-prone when derived directly from the observed choices. Hence, in the end, it may be necessary to estimate the three parameters jointly given the six choices of a person following, for instance, a maximum likelihood estimation approach. Section III.A outlines the motivation and details regarding the estimation of the risk parameters that can be obtained

from BAR.

A. Choice models for describing people's risk preferences

CPT-based Models. — For developing the BAR measure, we followed the assumptions of CPT. However, it is an open question whether simpler models such as a simplified expected utility theory (Von Neumann and Morgenstern, 2007) might be sufficient to describe people's risk preferences. Therefore, in addition to a three-parameter CPT-version (without a probability-weighting component, because we did not vary probabilities) we further test four additional, simplified models to describe the behavior for the BAR measure that are summarized in Table 1. First, we consider *CPT-1* entailing only one free parameter, assuming the same utility function for gains and losses. This model represents a simple expected utility model. Second, we include *CPT-2a* with two parameters, with separate utility function for gains and losses but no loss-aversion parameter. The third model – *CPT-2b* – assumes the same utility functions for gains and losses but accounts for the loss-aversion parameter. Fourth, *CPT-3* has three free parameters that account for different utility functions for gains and losses and loss aversion. This model corresponds to the full CPT model with no probability weighting. Finally, we will test a model – the *Expected Shortfall Model* (cf. ES) – that defines the utility of an option as the expected value of an option minus the potential shortfall. The shortfall of an option is defined by how much the lowest outcome of that option falls below a certain expectation or aspiration of a person. This different definition of utility distinguishes the ES model from the expected utility theory and the CPT.

By fitting these different models, we will be able to test which of the different components of the models are essential to describe people's risk preferences. More specifically, we test whether a distinction of people's risk preferences between the loss and gain domain is important. Furthermore, we investigate whether the assumption of loss aversion is essential to describe people's preferences. Researchers

argued that loss aversion could be seen as an orthogonal component of people's risk preferences (Köbberling and Wakker, 2005). Also, by testing two models that do not incorporate loss aversion (CPT 1 and 2a) against two models that do incorporate loss aversion (CPT 2b and 3), we can evaluate whether loss aversion is a crucial factor to explain behavior. Finally, the last model, the expected shortfall model, assumes that negative deviations of expectation or aspiration define the risk component of an option and can thus be used as a benchmark.

Finally, the testing of various models that do and do not account for the loss aversion allows us to test whether loss aversion is correlated with self-reported real-life risk-taking behavior and to investigate its retest reliability. In addition to examining the predictive accuracy of the different models, we also test the retest reliability of all individual risk parameters of these models. If the model parameters estimated from the data collected at two different time points lead to the same conclusions about a person's risk preference and these parameters are similarly correlated with real-life risk-taking, we can infer that an EM is robust. Also, apart from the standard expected utility models (i.e., CPT), we include a risk-value model (i.e., the ES model) stemming from the finance domain. According to this model, the expected utility of an option is a trade-off between the option's return and its risk (Sarin and Weber, 1993). Importantly, the ES model defines risk as expected shortfall (Acerbi and Tasche, 2002), rather than the variance of the choice outcomes, thus this model also focuses on the potential losses of an option.

Expected Shortfall Model. — The model assumes an alternative definition of risk focusing on losses. Despite its popularity in the finance area, it has rarely been used to estimate people's risk preferences (but see Weber et al., 2002, for an application of the expected shortfall). The model measures how much decision makers fall short in their expectations about the possible gain from option A with I outcomes x_i and the probability p_i. The shortfall of a choice option X is defined

as

$$Shortfall = E_{c,X} = \sum_{i=1}^{I} p_i[max(c - x_i, 0)], \tag{2}$$

where parameter c is an individual threshold, below which outcomes are undesired and considered as "losses". The shortfall is then subtracted from the expected value of the option (see formal description below). For the BAR measure, we used as a threshold for the loss and the gain items the outcome of the first option where both outcomes are the same. For the mixed gambles, we used zero as the threshold. Thus, the higher the negative deviation from the threshold, the higher the risk. Lopes and Oden (1999) showed that people tend to avoid the outcomes of their choices that fall short below their minimum expectation or "aspiration level".

Table 1—Possible Models Defining Utility in the Basel Risk Attitude Measure

Model	Parameters	Utility Function
CPT-1	α	$u(x) = x^{\alpha}$
CPT-2a	α, β	$u(x) = \begin{cases} x^{\alpha} & \text{, for } x \geq 0 \\ -x^{\beta} & \text{, for } x < 0 \end{cases}$
CPT-2b	$\alpha = \beta, \lambda$	$u(x) = \begin{cases} x^{\alpha} & \text{, for } x \geq 0 \\ -\lambda \times -x^{\alpha} & \text{, for } x < 0 \end{cases}$
CPT-3	α, β, λ	$u(x) = \begin{cases} x^{\alpha} & \text{, for } x \geq 0 \\ -\lambda \times -x^{\beta} & \text{, for } x < 0 \end{cases}$
ES	ω_{ES}	$U(X) = E[X] - \omega \times Shortfall$

Note: CPT = cumulative prospect theory models with 1, 2 or 3 free parameters. ES = Expected shortfall model. For all CPT-models, the expected utility of a choice option X is defined as $U(X) = \Sigma_i^I p_i u(x_i)$, which in case of the BAR task reduces to $U(X) = \Sigma_{i=1}^2 (x_i)/2$.

IV. Experimental Procedures

The experiment consisted of two sessions, each one month apart. In each session, we used the same EMs and questionnaires to replicate the results and to allow for estimating the measure's retest correlation.

A. Participants

Participants were recruited through Mechanical Turk (cf. mTurk). We determined the sample size of the study with a power analysis based on the results of a pilot study. [6] Effect sizes for retest reliabilities in the pilot study were around $r = 0.3$. For the retest reliability test, we aimed for a power of .80 to detect a retest-reliability correlation between the different measures of at least $r = 0.30$ with a probability of .80. Accordingly, a sample size of N=140 was required. Furthermore, we anticipated a dropout rate of around 30% from the first to the second experimental session, resulting in a total sample size of N=200 for the first session ($M_{age} = 36.7$, $SD_{age} = 9.7$, 43.5% female). After 31 days, 91% of the participants in session 1 ($N = 181$) signed up for the second session ($M_{age} = 37.5$, $SD_{age} = 9.8$, 44.2% female). Overall, 27% of participants reported a monthly income of below \$1000, 59% between \$1000 and \$5000 and 13% above \$5000. The professional status of participants included student (2%), employed (65%), self-employed (24.5%) and unemployed (8.5%). Of the participants, 8% reported working or studying in the financial domain. Participants were required to have not previously participated in earlier pilot versions of the study, to live in the US, and to have a success-rate for previous studies on mTurk of at least 90%.

B. Procedure

The study was approved by the institutional review board (IRB) of the Department of Psychology of the University of Basel . Participants were redirected

[6] We ran a pilot study in which people were asked to complete HL, EG, GP, SS, IMP scales and an early version of BAR at two time points, one month apart.

from mTurk to Unipark, a scientific questionnaire platform. During the recruitment procedure, they were informed about the payment (5$ for completion of each of the two sessions of the experiment plus a possible bonus of up to 6$ for each of the two sessions of the experiment). All participants supplied informed consent and were informed about the experiment in detail. They then completed the six different risk-preference tasks (HL, EG, GP, BAR, SS, IMP) in randomized order and the eight single real-life risk-taking questions (SOEP), followed by a demographic questionnaire asking about gender, age, income, profession, and whether they had ever worked or studied in the financial domain. Section A.A3 in Appendix A explains the Mechanical Turk procedure in detail.

In HL, GP, and EG, the lotteries chosen by each participant were played out using a random number generator and the outcome of the lottery was translated to the payment in USD. For every task, the range of outcomes of experimental currency converted to a range of 0-2 USD and the final payment was summed over all tasks. The bonus paid out was \$2.9 on average ($SD = \$.7$). Participants were reminded of session 2 during the bonus payment for session 1. Session 2 was initiated one month (31 days) after completion of session 1. Both sessions of the study had the same procedure, except that in session 1, participants were informed about the second session at the beginning and the end of the experiment. Participants were also informed that they would receive the monetary reward of session 1 regardless of whether they completed session 2.

C. Materials

Existing Risk-Preference Elicitation Measures. — We tested the Holt & Laury task (HL) using the common version by Holt et al. (2002) comprising 10 questions, with the last question offering a stochastically dominant option A over option B. We used this last option to check participants' understanding of the task. As a raw choice measure, we report the number of risky options that a participant took.

We used the original version of the Gneezy and Potters task (GP, Gneezy and Potters, 1997) with a hypothetical endowment of \$100, which was later translated to \$2 for the bonus payment. Participants could invest any amount of this experimental endowment in a lottery. The lottery resulted in a multiplication of the invested amount by 2.5 with a probability of 50%. Otherwise, the invested money was lost. The money that was not invested could be kept. We used the amount invested as the raw risk-preference measure. We tested the Eckel and Grossman task (EG) with the version in the original paper (Eckel and Grossman, 2002, see Table **??**, the Appendix A provides the exact lotteries and outcomes).

We measured sensation seeking (SS) with the Zuckerman Sensation Seeking Scale V by (Zuckerman, 2007) and impulsiveness (IMP) using the Barratt Impulsivity Questionnaire by Barratt (1965). These measures could not be incentivized.

Real-Life Risk Questions. — The real-life risk questions were translated from German into English from the German Socio-Economic Panel (SOEP). In these questions, participants were asked to rate their willingness in real-life to take risks in each of the following areas: risk-taking in general (SOEP-G), in driving, in financial matters, during leisure and sport, in occupational affairs, in health issues, in faith in other people and investment issues. For the main analyses, we used the question referring to the willingness to take risk in general as a benchmark for real-life risk-taking, because previous studies (i.e. Dohmen et al., 2011; Richter and Schupp, 2012) found that the SOEP-G is the best predictor of real-life risk-taking. The other SOEP questions all correlated with the general risk measures and were used as control measures. Detailed results regarding other SOEP domains are provided in Appendix B.

Basel Risk-Preference Measure (BAR). — Each of the six BAR questions was presented separately on one web-page and in a randomized order. As presented in Figure 1, the text of the question was presented on top, and the 14 options

were presented along with a horizontal slider, with Outcome 1 displayed above the slider and Outcome 2 below the slider. To display Outcomes 1 and 2 of each of the 14 choice options, participants had to click on the particular choice option on the slider, while outcomes of other choice options were invisible. A choice option could be chosen by selecting the choice option on the slider and clicking on "Proceed".

The participants were informed that they would receive a starting value of 1 USD. After the experiment was over, for each participant, one of the six items was randomly selected and played out. The range of the experimental currency for that item was normalized to a range of 0-1 USD. The gain (or loss) resulting from playing the lottery would then be added (or subtracted) from the staring value of 1 USD. Thus, participants could gain a bonus of between 0-2 USD, which was dependent on their choices in the task.

For the retest reliability analysis and the correlation with the behavior of the other measures as well as SOEP general, we used the mean rank choices of the BAR. Therefore, for each participant, we averaged the rank of the choices (which was between 1-14) over all 6 questions, which results in a single risk preference score from BAR. Additionally, we estimated the number of utility models listed in Section III.A for each participant, to estimate the individual risk parameters, which we report in Section V.D.

V. Results

We first provide descriptive results for the different EMs at session 1 and session 2 separately. Next, we report the pairwise correlation among all EMs. Section V.B reports the retest reliability using Spearman correlations of the scores of each EM between session 1 and session 2. In section V.C, we report each EM's correlation with SOEP-G as a benchmark for the relation to the real-life risk-taking. Section V.D provides the results of the modeling analysis for the BAR to report the best-fitting model. Finally, Section V.E outlines the parameter values

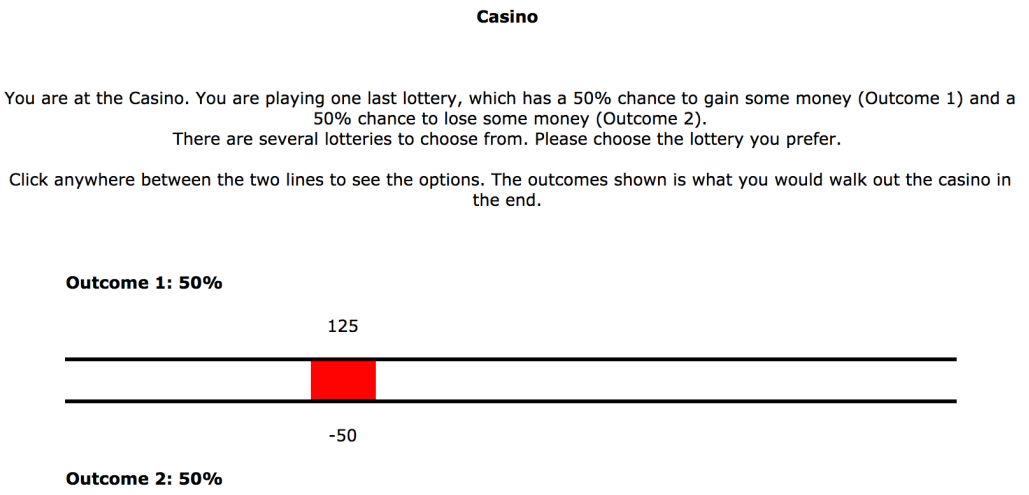

FIGURE 1. BAR QUESTION SCREENSHOT EXAMPLE

Note: Example question of the BAR task. Participants could click anywhere between the two lines to see the difference outcomes (positioned above and below the two slider-lines). After participants made a choice they could go to the next question via a continue button below.

of the best-fitting BAR models regarding retest reliability and correlation with SOEP-G. The parameters of the best-fitting BAR model are compared to the estimated parameters of the other EMs.

For the descriptive results, the retest reliability and the real-life validity (i.e. correlation with SOEP-G), we report the scores of each EM. The scores for the particular measures were the following: for the two personality measures SS and IMP, we used the total score (sum of all responses for each of the two measures); for the HL task, we used the number of risky options chosen; for the EG measure we report the choice of the gamble (1-6); and for the GP measure, the amount invested. For the BAR, we averaged the choices from all of the six items for each participant.

For the comparison of the risk parameters of the lottery EMs with the BAR parameter of the best-fitting models, we mapped the choices in the lottery EM

tasks (HL, EG, GP) onto risk-aversion parameters within a theoretical framework as proposed by Crosetto and Filippin (2016).

A. Descriptive Results

Table 2 shows means and standard deviations of the scores of all measures for session 1 and 2 separately. The mean scores and their standard deviations do not differ significantly between the sessions, according to paired t-tests and Levene tests. This shows that on average people's risk preferences did not change over the one-month period. The descriptive values are in line with the previous literature (Crosetto and Filippin, 2016; Frey et al., ress; Pedroni et al., ress).

Table 2—Descriptive results

	Session 1			Session 2		
	Mean	Median	SD	Mean	Median	SD
BAR	6.56	6.50	3.15	6.33	5.83	3.20
HL	3.79	4.00	1.99	4.00	4.00	2.17
GP	41.99	50.00	31.49	40.75	40.00	30.90
EG	3.15	3.00	1.56	3.46	3.00	1.51
SS	55.95	55.50	8.04	54.98	54.00	7.96
IMP	55.69	54.00	11.19	55.46	54.00	11.00
SOEP-G	5.07	4.50	2.74	4.98	5.00	2.53

Note: SS = sensation seeking. IMP = impulsiveness. SOEP-G = General willingness to take risk. BAR = Basel Risk Questions. GP = Gneezy & Potters Method. EG = Eckel & Grossmann Method. HL = Holt & Laury Method, SD = Standard Deviation.

In both sessions, all lottery-based measures (i.e., HL, EG, GP, BAR) correlated significantly with each other (r between 0.26 and 0.56). The self-report personality measures correlated lower with the lottery-based measures. SS correlated significantly with all EMs (r between 0.16 and 0.26), while IMP correlated weakly only with GP, EG, and BAR in session 1, and only with BAR in session 2. IMP did not correlate with HL in either session. The two self-report measures SS and IMP were significantly correlated with each other in both sessions. These results highlight two crucial findings. First, the BAR is significantly correlated to all risk measures in the set. Second, measures of the same type (self-report vs. lottery-

based) correlate more strongly with each other, which is in line with findings of Frey et al. (ress). Both session 1 and session 2 show similar correlations among the measures. Table B2 (session 2: B3) in Appendix B provides exact values of the correlations among the EMs' scores for sessions 1 and 2. The appendix also provides figures of distributions of raw choices for both session 1 B1 and session 2 B2.

B. Test-Retest Reliability

We evaluated the test-retest reliability with pairwise Spearman correlations between session 1 and session 2 for each EM. Also, to evaluate stability of our benchmark measure of the real-life risk-taking, we included the SOEP-G measure in the retest analysis. Retest reliability correlations for all measures were all positive and significant. The two personality questionnaire measures SS and IMP showed the highest retest correlation (SS: $r = 0.91$, $p < 0.001$, IMP: $r = 0.86$, $p < 0.001$). Second best retest correlation was shown by the BAR (BAR: $r = 0.61$, $p < 0.001$), followed by GP($r = 0.43$, $p < 0.001$), EG ($r = 0.39$, $p < 0.001$) and HL ($r = 0.32$, $p < 0.001$). When inconsistent participants (defined by only switching once from the safe to the risky options) were excluded from the HL analysis ($N = 6$), the retest measure for the HL measure improved slightly in retest correlation ($r = 0.36$, $p < 0.001$). Retest correlations after excluding these "inconsistent" participants are depicted Figure 2

C. External Validity

As a benchmark for the relation of each EM with real-life risk-taking behavior, we report the Spearman correlation for each EM with the SOEP-G. As shown in Figure 3, all measures were significantly correlated with the SOEP-G in both sessions. SS showed the highest correlation (SS: $r = 0.6$, $p < 0.001$). The BAR score correlated second best ($r = 0.43$, $p < 0.001$), followed by GP ($r = 0.41$, $p < 0.001$), EG ($r = 0.39$, $p < 0.001$), IMP ($r = 0.26$, $p < 0.001$) and HL

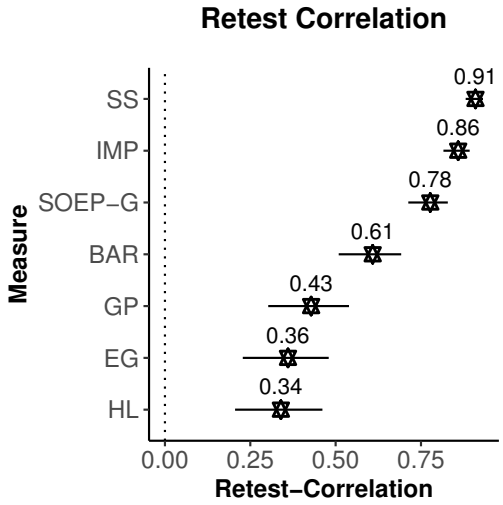

FIGURE 2. RETEST RELIABILITY

Note: Retest Correlation of the risk-preference measures. SS = sensation seeking. IMP = impulsiveness. SOEP-G = General willingness to take risk. BAR = Basel Risk Questions. GP = Gneezy & Potters Method. EG = Eckel & Grossmann Method. HL = Holt & Laury Method. Point ranges represent 95% CIs. $p < .001$ for all reported correlations.

($r = 0.2$, $p = 0.001$). When the six inconsistent participants were excluded from the HL, the HL measure improved slightly in retest correlation ($\delta r = 0.02$, $p < 0.001$). The larger retest reliability and external validity results for the BAR indicate that the measurement could be a more natural measure of people's risk preferences. Sensation seeking shows the highest correlation with the SOEP-G, whereas the correlation for the IMP measure is relatively low.

D. BAR Modeling Results

To test the importance of risk aversion versus loss aversion when considering people's risk preferences, we estimated the five theoretical utility models listed in Table 1 to the individual choices in BAR. We used maximum likelihood estimation (MLE) procedures to estimate the models' parameters, and we used the Bayesian Information Criterion (BIC) to compare the models against each other. The BIC takes the fit of the model and the model's complexity into account (see

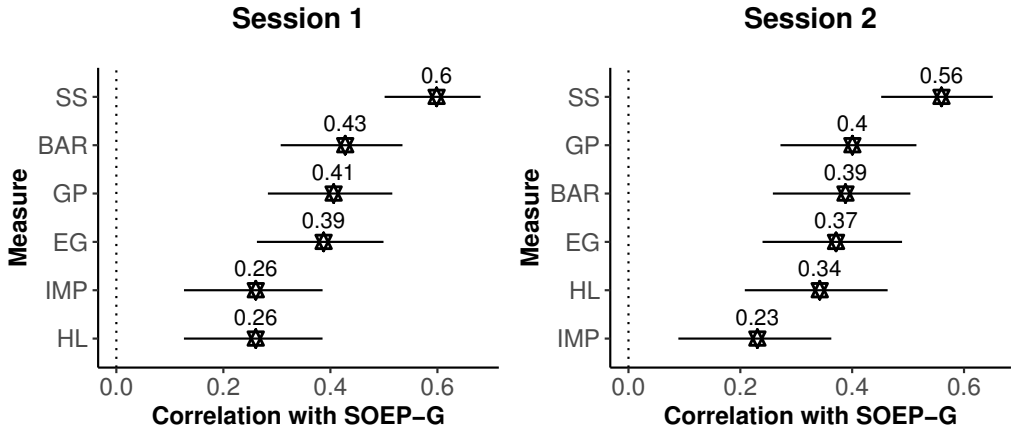

FIGURE 3. SPEARMAN CORRELATIONS BETWEEN EMS AND GENERAL RISK-TAKING BEHAVIOR

Note: Left Panel: Correlation of all EMs with SOEP-G in session 1 Right Panel: Correlation of all EMs with SOEP-G in session 2. SS = sensation seeking. IMP = impulsiveness. BAR = Basel Risk Attitude Questions rank mean. GP = Gneezy & Potters Method. EG = Eckel & Grossmann Method. HL = Holt & Laury Method. Point ranges represent 95% CIs. $p < .005$ for the HL correlation in session 1 and IMP correlation in session 2. All other correlations were significant at $p < .001$.

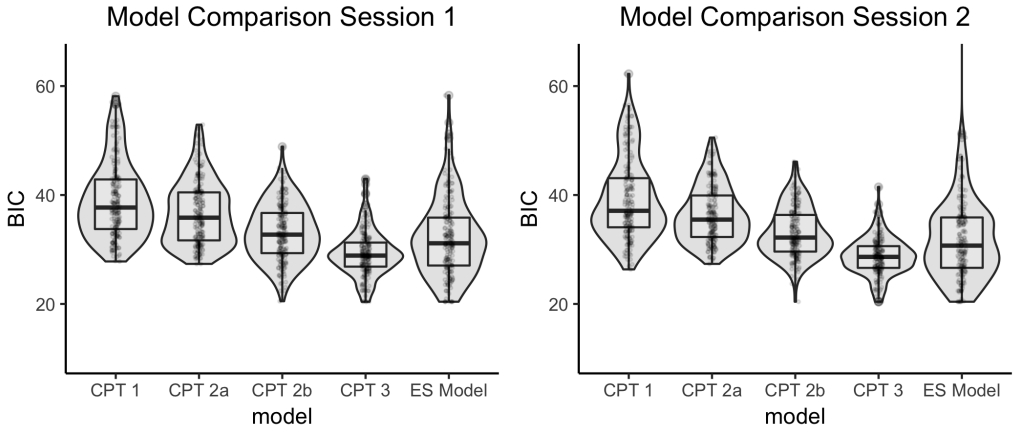

FIGURE 4. MODEL COMPARISON

Note: Bean plots of individual BIC values for each of the calculated models for the BAR in session 1 (left panel) and session 2 (right panel).

Lewandowsky and Farrell, 2010). Section A.A4 in Appendix A describes the model fitting procedures in more detail.

Figure 4 shows the distributions of BIC values for the five models in session

Table 3—BAR Model Comparison

	Session 1			Session 2		
Model	LL1	BIC1	No.bestfit1	LL2	BIC2	No.bestfit2
CPT-1	-17.67	38.94	7	-17.78	39.12	1
CPT-2a	-16.48	36.55	0	-16.38	36.32	0
CPT-2b	-14.75	33.10	10	-14.79	33.14	9
CPT-3	-12.80	29.19	128	-12.56	28.69	129
ES	-14.22	32.04	55	-14.08	31.71	46

Note: CPT = Cumulative prospect theory model with 1, 2 or 3 free parameters. *ES Model* = Expected shortfall model. *LL* = Average log-likelihood. *BIC* = Average BIC values. Smaller values for BIC and larger values for the LL represent better fit. *No.bestfit* = Number of people for which the respective model best described the data compared to the other models, based on the BIC.

1 (left panel) and session 2 (right panel), while Table 3 outlines the models' fit information. According to mean BIC in both sessions, a number of participants for whom the model described the data best and the distribution with the least variance of the BIC values, CPT-3 described the data best. CPT-2b has the second lowest mean BIC, which was slightly better than the BIC values for the ES model. However, when comparing the models for each participant the ES model accounted best for more participants than CPT-2b. The CPT-2a and CPT-1 models describe the data least accurately, and only accounted best for the choices of five participants altogether, based on their BIC values.

About two-thirds of the participants were best described by the CPT-3 model in both sessions. The expected shortfall model accounted best for another third of the participants. Models that did not account for loss aversion (CPT-1 and CPT-2a) were selected as the best model for only seven (session 1) and one (session 2) participants together.

As an interim conclusion, these findings support the importance of accounting for losses and loss aversion in measuring people's risk preferences. The fact that CPT-3, CPT-2b and the ES model outperform the other models indicates that loss aversion plays an important role in predicting people's risk preferences. Despite indicating the importance of separating the utility functions for gains and losses, our results highlight the great importance of accounting for loss aversion. This

is seen by the better performance of all CPT versions that include λ and the ES model, over the models that do not account for the loss aversion.

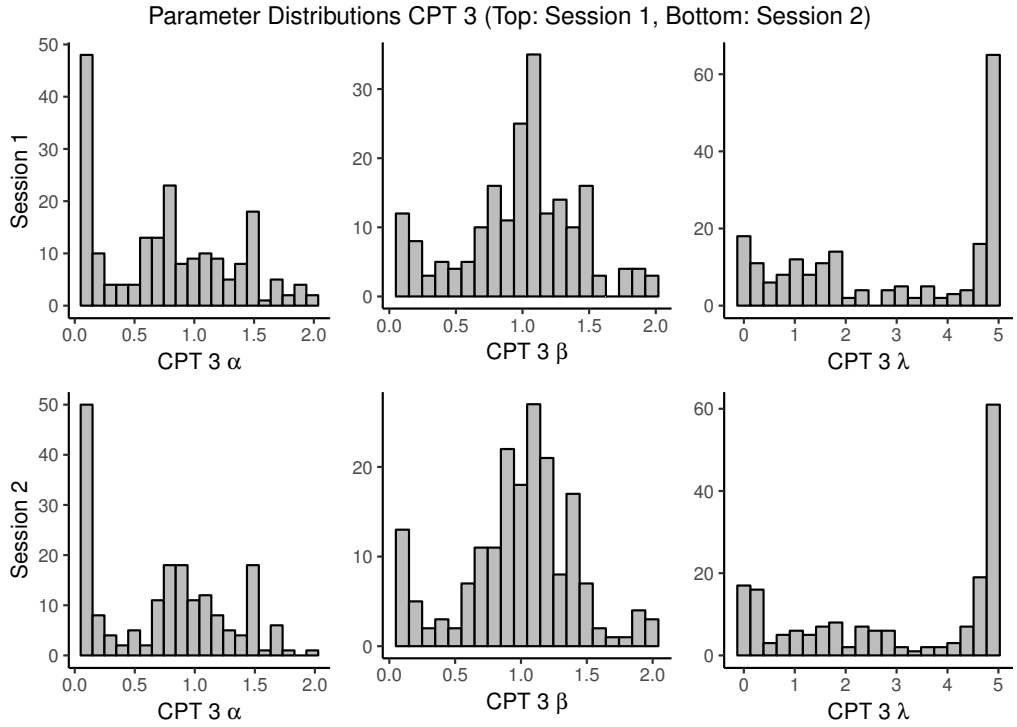

Figure 5. Distributions of Parameter Values of CPT-3

Note: Distributions of estimated parameter values for CPT-3 in both sessions. Top: Session 1, Bottom: Session 2.

Parameter Estimates. — Figure 5 shows the distributions of the three parameters of the best-fitting model (CPT-3). Recall that α defines the curvature of the utility function for gains, β defines the curvature of the utility function for losses, and λ defines how people trade off gains to losses (i.e., loss aversion). For α and β parameters, a value of 1 represents risk neutrality. For the gain domain, values lower than 1 represent risk aversion; however, for the loss domain, values lower than 1 represent risk-seeking preference (and values *above* 1 represent risk

aversion). According to the findings for cumulative prospect theory of (Tversky and Kahneman, 1992), both α and β fall around a value of .88 and λ around 2.25.

The distribution of estimated α parameter was positively skewed, and the distribution of λ was negatively skewed. Thus, CPT-3 predicts a large number of participants as highly risk-averse for gains and loss averse, which is in agreement with the assumptions of the prospect theory. Even though we obtained α of close to 0 for almost a quarter of our participants, the median α was .74 and.76 in session 1 and 2, respectively. The distribution of β was close to normal. However, the median of the estimated β values in session 1 and session 2 were 1.04 and 1.02 consecutively, which indicates risk aversion rather than risk-seeking also for losses, contrary to the assumptions of prospect theory. In both sessions, the median loss aversion λ equaled 3.17 and 3.68, which is higher than what was found by (Tversky and Kahneman, 1992). Additionally, for around 40 participants, the parameter estimation yielded λ values of more than 4.9 and because we set the boundaries of the estimation for λ to 5, these values could likely be higher if the boundaries were set higher. These findings indicate that people were loss averse as assumed by prospect theory and showed stronger loss aversion than in previous studies.

The unusual finding that high risk aversion for losses and very high loss aversion are directly linked can be problematic when estimating the parameters over all six questions of the BAR. Participants with high estimated λ values made risk-averse choices for the loss items of the BAR (resulting in high β values) as well as risk-averse choices for the gain items (resulting in low α values). Recall now that λ is defined in the mixed gambles and is dependent on alpha and beta. Next, a participant who is risk-averse in both gains and losses will be assumed to have high β and low α. If then λ is estimated in the mixed domains, losses are very high (because of the high exponent β) and gains are very low (because of the low exponent α). Thus, to compensate for these inequalities between gains and losses, to choose any of the gambles in the mixed domain, the participant will have to

be assumed to have a very high λ. Figure B4 in the appendix depicts this and shows the histograms of choices in the BAR for session 1, color-coded with gray for high (i.e., values for λ over 4) and black for low estimates of λ.

The estimated parameters of CPT-3 also correlated with each other. We observed that λ negatively correlated with α (session 1: $r = -0.47, p < .001$, session 2: $r = -0.52$, $p < .001$) and positively correlated with β (session 1: $r = 0.54, p < .001$, session 2: $r = 0.57$, $p < .001$). However, α moderately correlated with β only in session 2 ($r = -0.2$, $p = 0.004$) but not in session 1 ($r = 0.06$, $p = 0.394$). This could either mean that loss aversion is related to risk aversion in both gain and loss domains, but risk aversion in gains and losses may not necessarily relate, which supports the criticism leveled at EMs that do not incorporate losses. Another explanatory implication of this is that it is difficult to estimate the parameters of the CPT-3 model separately since they correlate so highly with each other. This hypothesis is supported by the explanation of very high λ values above.

Next, we investigated the parameter estimations of the second- and third-best models, ES and the CPT-2b. Recall that the risk parameter ω in the ES model represents a weight that is assigned to the choice outcomes that fall short of one's expectations; that is, they are cognitively treated as losses. Therefore, the larger the ω, the higher is a person's sensitivity to losses. The ES model parameter ω peaks around zero and has a median of 0.12 (0.51), which indicates that most people are risk-averse, where risk is defined as the shortfall from an expectation level. For 33% of the participants, the shortfall parameter ω was between -0.1 and 0.1, indicating a very small weight on the shortfall compared to one's expectations, suggesting on average risk-neutrality. However, the majority of estimates were higher than 0 and thus the majority of people were risk-averse, based on the classification of the ES model.

CPT-2b has only one parameter for the utility function for gains and losses and also the loss-aversion parameter λ. The parameters of the CPT-2b model do not

show skewed distributions, and the median of the α is 0.87 (0.92), which indicates that people are risk averse for gains and risk seeking for losses. The λ parameter shows a median of 1.97 (2.15). Both parameter values are in line with previous estimations of these two values and also in line with previous findings of these estimations (Tversky and Kahneman, 1992). Although the model using these parameters would correctly predict the average behavior in the gain and mixed-outcome domain of the BAR (risk-averse choices for gains and loss-averse choices for mixed gambles), the model does not predict the majority of risk-averse choices in the loss domain, which explains the lower fit of the CPT-2b model compared to the CPT-3 model. This speaks of having two separate utility functions for gains and losses. Distributions of the estimated parameters of the ES and CPT-2b model are presented in figure B3 of the appendix B.

CPT-3 WITH DERIVED PARAMETERS. — Instead of following the maximum likelihood approach, the parameter values of the CPT-3 model can also be derived directly from the choices of participants, at least for the gain and loss items, by taking the pre-calculated estimates for each parameter (Tables A.A1 - A4 of Appendix A show the parameter values that correspond to the particular choice options in BAR). To get an estimate for α and β parameters we averaged the two gain and loss questions, respectively. As a reference value, we took the midpoint of the respective ranges of parameters for each option. As pointed out above for the items with mixed outcomes (i.e., losses and gains), the parameter values of λ would only hold if people had a specific parameter value for α and β (i.e., $\alpha = \beta = .9$). However, because most likely people's choices deviated from these values, the derived values for λ only represent an approximation. Thus the fit of the CPT-3 derived procedure will necessarily be less good than the for when CPT-3 parameters are optimized.

We compared CPT-3 with derived and optimized parameter values against each other by correlating the parameters with each other. All three derived parameters

correlated highly with the respective CPT-3 model parameter that were optimized – α: $r = 0.93$, $p < 0.001$, β: $r = 0.85$, $p < 0.001$, λ: $r = 0.85$, $p < 0.00$. This indicated that reading out the parameters from the pre-calculated tables (see B4 in appendix B) serves as a good approximation of the estimates calculated by an MLE procedure. Paired sample t-tests revealed that the means of α ($\Delta M = 0.02$, 95% CI [−0.12, 0.07], $t(361.89) = -0.49$, $p = .622$) and β ($\Delta M = 0.05$, 95% CI [−0.13, 0.04], $t(394.54) = -1.10$, $p = .270$) parameters were not significantly different when CPT-3 was optimized than when it was derived. However, the λ parameter values were significantly lower when derived from the raw choices than when estimated with optimizing procedures ($\Delta M = -1.05$, 95% CI [0.77, 1.34], $t(261.26) = 7.24$, $p < .001$). This effect can easily be explained because the optimizing procedure of λ allowed the values to go up to 5 and due to the unusual loss-averse choices in the loss domain (see an elaboration of this issue above).

Finally, we tested how many participants were categorized as risk-averse or risk-seeking by both procedures, CPT-3 with derived and optimized parameter values. Based on the α parameter, both CPT-3 procedures led to the same categorizations for 94% of the participants, based on the β parameter for 90% of the participants and based on the λ parameter for 91.5% of the participants. Thus, we conclude that deriving the individual risk parameters from the pre-calculated parameter values does approximate the parameters optimized by maximum likelihood procedures sufficiently . This result enables the BAR to be used as a categorization tool without estimating the exact parameters with an optimizing procedure.

Retest Reliability and Correlation with SOEP-G for CPT-3 and ES. — We tested the test-retest reliability and correlation with the self-reported real-life risk-taking (i.e., SOEP-G score) for the three optimized parameters of CPT-3 and the optimized parameter of the ES model. For each of the four parameters, we calculated a Spearman correlation between session 1 and session 2, and correla-

tions for each parameter in each session with the SOEP-G score. As shown in Table 4, the ES parameter had the highest retest reliability of all the parameters and also correlated the best with the SOEP-G score in both sessions. However, the CPT-3 λ parameter showed the highest retest reliability out of its three parameters and was not substantially worse than the ES ω parameter. Altogether this shows that loss aversion and the attention to losses as described by the shortfall model are the most robust criteria for eliciting risk preferences. However, α (i.e., risk aversion in the domain of gains) has higher retest reliability than β as well as correlation with the SOEP-G score.

Table 4—Test Criteria for Model Parameters

	Retest	Correlation with SOEP-G	
	Correlation	Session 1	Session 2
CPT-3 α	0.4 (<0.001)	0.22 (0.005)	0.24 (<0.001)
CPT-3 β	0.24 (<0.001)	-0.22 (0.005)	-0.22 (0.005)
CPT-3 λ	0.44 (<0.001)	-0.33 (<0.001)	-0.34 (<0.001)
ES ω	0.57 (<0.001)	-0.39 (<0.001)	-0.36 (<0.001)

Note: Test criteria for the estimated model parameters of the BAR. *CPT-3*= Cumulative prospect theory model with three free parameters for gains, losses, and mixed gambles.*ES* = Expected shortfall model with one parameter.

E. Comparison of Risk-Preference Estimates from Different Models

We compared the parameter estimates of the two best-fitting models of the BAR (CPT-3 and ES) with the risk parameters estimated by the other EMs - GP, HL, and EG. From the CPT-3 we only show the correlations with the α parameter because the α value is based on only gain questions such as HL, GP and EG measures. For better comparability, we recalculated the original risk parameter estimates of the HL, GP and EG measures to a common utility function in the form of x^r (see Crosetto and Filippin, 2016, for an elaboration). Even though the ES model parameter has a different definition of utility, we included it here to compare the correlation with the other measures for explorative reasons. Table 5

shows the correlations among the risk parameter estimates given by the different measures (or models) including the parameter value α of the CPT-3 model and the ES Model parameter ω. All risk parameters are positively[7] and significantly correlated. The CPT-3 α parameter correlated substantially with the GP measure and moderately with the EG and the HL measure. These results were about the same for the ES parameter. These results indicate that the measures all seem to measure, to a certain extent, a similar construct of risk-preference; however, the fact that those measures aim to measure the same underlying construct diminished the substantiality of this result. Similar low convergent cross-validities were also reported in recent large-scale findings (Pedroni et al., ress).

Table 5—Correlation Risk Parameters Assuming the Power-Utility Function

	HL	EG	GP	BAR α
HL				
EG	0.33(<.001)			
GP	0.23(0.002)	0.21(0.003)		
CPT-3 α	0.24(0.001)	0.23(0.001)	0.40(<.001)	
ES ω	-0.24(0.001)	-0.25(<.001)	-0.38(<.001)	-0.61(<.001)

Note: *HL* = Holt and Laury measure. *EG* = Eckel and Grossmann measure. *GP* = Gneezy and Potters measure. *CPT-3 & ES* = parameters of different BAR measures.

VI. Discussion and Conclusion

In scientific investigation and practical applications, it is crucial to measure people's risk attitude; however, this task is not trivial. Over recent years, a large battery of various measures has been developed (see Frey et al., ress, for a review). The main challenge in selecting the best measure for quickly estimating people's risk appetite in investment decisions is the trade-off between short elicitation time, consistency of the measure across time, and obtaining a quantitative output of the measure. On the one hand, self-reported measures are more consistent across

[7] ω parameter of the ES model is negative because lower values represent higher risk aversion, which is the opposite for the other measures

time. On the other hand, lottery-based methods provide a quantitative output that can be treated as an objective *risk parameter*.

In the present work, we identify challenges and the most desirable features of various existing risk measures. We propose BAR – a new method by combining a behavioral approach using lotteries with a self-report measurement approach that simulates real-life scenarios. This new method provides a quantitative output in the form of a utility function for gains and losses as well as a measure of loss aversion, following the line of thinking in cumulative prospect theory. This feature enables the complete spectrum of one's risk preferences to be captured, which distinguishes the measure from existing EMs.

Also, the new measure correlates with self-reported real-life risk-taking in the form of a single willingness-to-take-risk question, which addresses the problem of the ecological validity of the risk-preference EM. The BAR also takes no more than five minutes to complete, which seems to be a requirement for many practitioners using risk-preference elicitation measures in financial contexts. We contrasted this measure with the existing most popular lottery-based and survey-based risk measures. Also, we evaluated the test-retest reliability of all risk measures included in the study.

First, we found that all measures positively correlate with each other, such that in general and relative to the rest of the sample, a more risk-seeking person classified by one measure would also be classified as more risk-seeking on another measure. In our sample, we find that the majority of participants were risk-averse and loss-averse. Second, all measures correlate positively with a benchmark for real-life risk-taking (SOEP-G); however, the IMP and HL measures correlate only weakly.

Our results indicate that self-reported risk attitude measures such as personality measures offer higher retest reliability. This may result from the fact that participants may have an image of self and act accordingly when responding to verbalized questions, whereas lottery-based questions present abstract non-

verbalized choices without context. Also, the self-reported measures may be easier to comprehend than numerical values (Dave et al., 2010). The new BAR measure performs substantially better than the lottery-based methods on both the test-retest reliability and external validity, measured on a benchmark of the SOEP-G question. BAR's performance on both dimensions is comparable to the self-reported measures such as sensation seeking, which supports that the new measure has a survey-like structure and most likely is easier to comprehend than lotteries.

One of the main advantages that the BAR offers is its distinction between gains and losses and the estimation of a loss-aversion parameter because a major challenge of risk-preference measurement is the prevailing difference of definitions of risk. Practitioners and laypeople define risk as the possibility of losses, whereas finance and decision-making researchers often define it as the variance of possible outcomes (Schonberg et al., 2011). This difference in definitions might lead to a problem when using measures that only follow the variance definition and do not incorporate the domain of losses or distinguish between losses and gains. In our measure, we introduce these valence domains. Also, our analyses show that when loss aversion is estimated separately from utilities in the gain and loss domains, its retest reliability as well as its correlation with real-life risk-taking improves, compared to the other parameters or the parameters of other EMs.

Finally, our modeling analysis shows that risk-taking behavior in the BAR can be best described by a three-parameter model that is akin to cumulative prospect theory models (Tversky and Kahneman, 1992). The model explained the data better than other models that were in line with CPT and only had two or fewer parameters. We could also show that parameters derived from the raw choices by reading out the pre-calculated parameter values from a table resulted in a good approximation of the optimized parameter values of the CPT-3 model. Finally, a different but simpler one-parameter model that defined the utility in an option based on the expected shortfall of a certain expectation fit the data of

participants almost as well as the best-fitting CPT-3 model. Most importantly, the risk parameter of the shortfall model provided a better retest reliability as well as a better correlation with SOEP-G. To conclude, it is not resolved which model should be used as a basis for the risk parameter elicited with the BAR. If someone prefers three separate risk preference values for gains, losses, and loss aversion it is appropriate to read out three different parameters. However, the ω parameter of the shortfall (ES) model can be used as an alternative more robust estimate, which would be in line with the Occam's razor debate.

Despite several novel contributions, this study faced a number of challenges. First, when estimating multiple parameters over only six questions, the modeling procedures often do not converge or they result in extreme parameter estimates. Additionally, risk parameters within one model were often highly correlated, which could indicate an issue when aiming to estimate separate psychological factors. A possibile way around this would be to use hierarchical Bayesian parameter estimation techniques that account for large distributions from a population mean (Nilsson et al., 2011).

Further, a large proportion of participants often chose extreme values in the BAR, which indicates that the options might not capture the full range of preferences (see choice distributions in the appendix figure B1 and B2). A large proportion of people was either very risk averse or very risk seeking. Even though the former effect is common among tasks that restrict the answer options, this needs to be addressed in further studies when improving the measure. A possible extension would include fewer choice options in the middle of the range of α and β and more choice options for the extremely risk-seeking and risk-averse participant. This approach would possibly answer the question of whether estimated parameters reached extreme values (i.e., allowed boundaries) due to inconsistent choices or insufficient data, or because the range of the allowable choice options in the risk-seeking domain was too small.

In sum, our findings have multiple implications. First, current short measures

of risk preference can be improved by multiple simple changes such as the introduction of contexts, keeping the measure simple to answer and understandable for participants and defining more fine-grained choice options to enhance the precision of the estimation of risk parameters. Second, our results indicate the importance of allowing loss aversion to be measured, which, to our knowledge, has not been done before with a simple elicitation method. We believe that the newly proposed BAR method can provide new directions in developing more accurate, reliable, and understandable risk measures. Also, we believe that the findings of this paper will draw readers' attention to important aspects in further advancements of risk-appetite elicitation.

References

Abdellaoui, M., A. Driouchi, and O. L'Haridon (2011). Risk aversion elicitation: Reconciling tractability and bias minimization. *Theory and Decision 71*(1), 63–80.

Acerbi, C. and D. Tasche (2002). On the coherence of expected shortfall. *Journal of Banking & Finance 26*(7), 1487–1503.

Barratt, E. (1965). Factor analysis of some psychometric measures of impulsiveness and anxiety. *Psychological Reports 16*(2), 547 –554.

Boskovic, T., C. Cerruti, and M. Noel (2010). *Comparing European and US Securities Regulations: MiFID versus Corresponding US Regulations*. Washington, DC: World Bank.

Bruhin, A., H. Fehr-Duda, and T. Epper (2010). Risk and Rationality: Uncovering Heterogeneity in Probability Distortion. *Econometrica 78*(4), 1375–1412.

Camerer, C. F. (1998). Prospect theory in the wild: Evidence from the field. *Social Science Working Paper 1037*.

Charness, G., U. Gneezy, and B. Halladay (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization 131*, 141–150. WOS:000387517900011.

Charness, G., U. Gneezy, and A. Imas (2012). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization 87*, 43–51.

Charness, G. and A. Viceisza (2012). Comprehension and Risk Elicitation in the Field: Evidence from Rural Senegal. *Departmental Working Papers*.

Crosetto, P. and A. Filippin (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics 19*(3), 613–641.

Dave, C., C. C. Eckel, C. A. Johnson, and C. Rojas (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty 41*(3), 219–243.

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

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

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

Eckel, C. C. and P. J. Grossman (2008b). Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results 1*, 1061–1073.

Frey, R., A. Pedroni, R. Mata, J. Rieskamp, and R. Hertwig (in press). Risk preference shares the structure of major psychological traits. *Science Advances*.

Galizzi, M. M., S. R. Machado, and R. Miniaci (2016). Temporal Stability, Cross-Validity, and External Validity of Risk Preferences Measures: Experimental Evidence from a UK Representative Sample. SSRN Scholarly Paper ID 2822613, Social Science Research Network, Rochester, NY.

Gneezy, U. and J. Potters (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics 112*(2), 631–645.

Hittner, J. B. and R. Swickert (2006). Sensation seeking and alcohol use: A meta-analytic review. *Addictive Behaviors 31*(8), 1383–1401.

Holt, C. A., S. K. Laury, and others (2002). Risk aversion and incentive effects. *American economic review 92*(5), 1644–1655.

Josef, A. K., D. Richter, G. R. Samanez-Larkin, G. G. Wagner, R. Hertwig, and R. Mata (2016). Stability and change in risk-taking propensity across the adult life span. *Journal of Personality and Social Psychology 111*(3), 430–450.

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

Kahneman, D. and A. Tversky (2000). *Choices, Values, and Frames*. Cambridge University Press.

Köbberling, V. and P. P. Wakker (2005). An index of loss aversion. *Journal of Economic Theory 122*(1), 119–131.

Lejuez, C. W., W. M. Aklin, H. A. Jones, J. B. Richards, D. R. Strong, C. W. Kahler, and J. P. Read (2003). The balloon analogue risk task (BART) differentiates smokers and nonsmokers. *Experimental and Clinical Psychopharmacology 11*(1), 26–33.

Lewandowsky, S. and S. Farrell (2010). *Computational Modeling in Cognition: Principles and Practice*. Sage Publications.

Lönnqvist, J.-E., M. Verkasalo, G. Walkowitz, and P. C. Wichardt (2015). Measuring individual risk attitudes in the lab: Task or ask? An empirical comparison. *Journal of Economic Behavior & Organization 119*, 254–266.

Lopes, L. L. and G. C. Oden (1999). The Role of Aspiration Level in Risky Choice: A Comparison of Cumulative Prospect Theory and SP/A Theory. *Journal of Mathematical Psychology 43*(2), 286–313.

March, J. G. and Z. Shapira (1987). Managerial Perspectives on Risk and Risk Taking. *Management Science 33*(11), 1404–1418.

Menkhoff, L. and S. Sakha (2014). Multiple-item risk measures. Technical Report 1980, Kiel Working Paper.

Mishra, S., M. L. Lalumière, and R. J. Williams (2010). Gambling as a form of risk-taking: Individual differences in personality, risk-accepting attitudes, and behavioral preferences for risk. *Personality and Individual Differences 49*(6), 616–621.

Mullen, K. M., D. Ardia, D. L. Gil, D. Windover, and J. Cline (2009). DEoptim: An R package for global optimization by differential evolution.

Murphy, R. O. and R. H. ten Brincke (2017). Hierarchical maximum likelihood parameter estimation for cumulative prospect theory: Improving the reliability of individual risk parameter estimates. *Management Science*.

Nilsson, H., J. Rieskamp, and E.-J. Wagenmakers (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology 55*(1), 84–93.

Pedroni, A., R. Frey, A. Bruhin, G. Dutilh, R. Hertwig, and J. Rieskamp (in press). The risk elicitation puzzle. *Nature Human Behavior*.

Penolazzi, B., P. Gremigni, and P. M. Russo (2012). Impulsivity and Reward Sensitivity differentially influence affective and deliberative risky decision making. *Personality and Individual Differences 53*(5), 655–659.

Pratt, J. W. and R. J. Zeckhauser (1987). Proper risk aversion. *Econometrica: Journal of the Econometric Society 55*(1), 143–154.

Richter, D. and J. Schupp (2012). SOEP Innovation Sample (SOEP-IS)—Description, Structure and Documentation. *SOEP Working Papers*.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition 34*(6), 1446–1465.

Sarin, R. K. and M. Weber (1993). Risk-value models. *European Journal of Operational Research 70*(2), 135–149.

Schonberg, T., C. R. Fox, and R. A. Poldrack (2011). Mind the gap: Bridging economic and naturalistic risk-taking with cognitive neuroscience. *Trends in Cognitive Sciences 15*(1), 11–19.

Tversky, A. and D. Kahneman (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty 5*(4), 297–323.

Von Neumann, J. and O. Morgenstern (2007). *Theory of Games and Economic Behavior*. Princeton, UK: Princeton university press.

Wagner, G. G., J. R. Frick, and J. Schupp (2007). The German Socio-Economic Panel Study (SOEP) - Evolution, Scope and Enhancements. SSRN Scholarly Paper ID 1028709, Social Science Research Network, Rochester, NY.

Wakker, P. P. (2010). *Prospect Theory: For Risk and Ambiguity*. UK: Cambridge university press.

Weber, E. U., A.-R. Blais, and N. E. Betz (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making 15*(4), 263–290.

Wilcox, R. R. (2011). *Introduction to Robust Estimation and Hypothesis Testing.* Academic press.

Wong, A. and B. J. Carducci (1991). Sensation seeking and financial risk taking in everyday money matters. *Journal of business and psychology 5*(4), 525–530.

Zaleśkiewicz, T. (2001). Beyond risk seeking and risk aversion: Personality and the dual nature of economic risk taking. *European Journal of Personality 15*(S1), S105–S122.

Zuckerman, M. (2007). Sensation Seeking and Risk. In *Sensation Seeking and Risky Behavior.*, pp. 51–72. Washington: American Psychological Association.

Appendix A: Detailed Procedures

Appendix A follows up on some of the fine-grained methodological details of the measures and provides more-detailed information about the statistical analysis, including modeling procedures. In section A.A1 we report the exact text of the BAR questions and the exact values of outcomes of each question. Section A.A2 explains and in the EG measure. Section A.A3 then describes detailed procedures for the participant acquisition and data collection on Mechanical Turk. Finally, section A.A4 describes in detail the methods used for the estimation of the BAR risk parameters.

A1. BAR Questions

Table A1—BAR Questions

No.	Question Text	Domain
1	You are at the Casino. Your current gambling balance is 140$. You can now play a lottery, which has a 50% chance to increase your gains (Outcome 1) and a 50% chance to lower your gains (Outcome 2). There are several lotteries to choose from. Please choose the lottery you prefer.	+ small
2	You have invested money in the stock market. Your current investment balance is 1120$. You can now make an investment, which has a 50% chance to increase your gains (Outcome 1) and a 50% chance to lower your gains (Outcome 2). There are different investment options available. Please choose the investment you prefer.	+ large
3	You are at the Casino. Your current gambling balance is -140$. You can now play a lottery, which has a 50% chance to reduce your losses (Outcome 1) and a 50% chance to increase your losses (Outcome 2). There are several lotteries to choose from. Please choose the lottery you prefer.	- small
4	You have invested money in the stock market. Your current investment balance is -1120$. You can now make an investment, which has a 50% chance to reduce your losses (Outcome 1) but a 50% chance to increases your losses (Outcome 2).	-large
5	You are at the Casino. You are playing one last lottery, which has a 50% chance to gain some money (Outcome 1) and a 50% chance to lose some money (Outcome 2). There are several lotteries to choose from. Please choose the lottery you prefer.	+/- small
6	You invest in the stock market. You are now making an investment, which has a 50% chance for gain (Outcome 1) and a 50% chance for a loss (Outcome 2). There are different investment options available. Please choose the investment you prefer.	+/- large

Note: The table shows the exact wording of questions in the BAR measure. Questions were presented separately each on one page and presented in a randomized order. Below each question was a slider which could be selected to see the outcome options.

Table A2—BAR Measure Option Values of Gain Domain

Context	Option	Outcome 1	Outcome 2	Low	Range	High	EV	SD
casino	1	140	140.00	0.00	$< \alpha \leq$	0.21	140.00	0.00
	2	130	150.60	0.21	$\leq \alpha \leq$	0.30	140.30	10.30
	3	120	162.30	0.30	$\leq \alpha \leq$	0.42	141.15	21.15
	4	110	174.80	0.42	$\leq \alpha \leq$	0.53	142.40	32.40
	5	100	187.70	0.53	$\leq \alpha \leq$	0.63	143.85	43.85
	6	90	200.70	0.63	$\leq \alpha \leq$	0.75	145.35	55.35
	7	80	213.20	0.75	$\leq \alpha \leq$	0.85	146.60	66.60
	8	70	225.00	0.85	$\leq \alpha \leq$	0.95	147.50	77.50
	9	60	235.70	0.95	$\leq \alpha \leq$	1.05	147.85	87.85
	10	50	245.00	1.05	$\leq \alpha \leq$	1.15	147.50	97.50
	11	40	252.80	1.15	$\leq \alpha \leq$	1.24	146.40	106.40
	12	30	259.00	1.24	$\leq \alpha \leq$	1.34	144.50	114.50
	13	20	263.50	1.34	$\leq \alpha \leq$	1.44	141.75	121.75
	14	10	266.30	1.44	$\leq \alpha \leq$	2.00	138.15	128.15
stock-market	1	1120	1120.00	0.00	$\leq \alpha \leq$	0.21	1120.00	0.00
	2	1040	1204.80	0.21	$\leq \alpha \leq$	0.30	1122.40	82.40
	3	960	1298.40	0.30	$\leq \alpha \leq$	0.42	1129.20	169.20
	4	880	1398.40	0.42	$\leq \alpha \leq$	0.53	1139.20	259.20
	5	800	1501.60	0.53	$\leq \alpha \leq$	0.63	1150.80	350.80
	6	720	1605.60	0.63	$\leq \alpha \leq$	0.75	1162.80	442.80
	7	640	1705.60	0.75	$\leq \alpha \leq$	0.85	1172.80	532.80
	8	560	1800.00	0.85	$\leq \alpha \leq$	0.95	1180.00	620.00
	9	480	1885.60	0.95	$\leq \alpha \leq$	1.05	1182.80	702.80
	10	400	1960.00	1.05	$\leq \alpha \leq$	1.15	1180.00	780.00
	11	320	2022.40	1.15	$\leq \alpha \leq$	1.24	1171.20	851.20
	12	240	2072.00	1.24	$\leq \alpha \leq$	1.34	1156.00	916.00
	13	160	2108.00	1.34	$\leq \alpha \leq$	1.44	1134.00	974.00
	14	80	2130.40	1.44	$\leq \alpha \leq$	2.00	1105.20	1025.20

Note: BAR choice options for the two gain questions. Range indicates the α-parameter range each option implied. Participants only saw columns "Outcome 1" and "Outcome 2". Information in columns "Option" (Option Number), "EV" (Expected Value) and "SD" (Standard Deviation) was not provided to participants.

Table A3—BAR Measure Option Values of Loss Domain

Context	Option	Outcome 1	Outcome 2	Low	Range	High	EV	SD
casino	1	-140	-140.00	1.43	$\leq \beta \leq$	2.00	-140.00	0.00
	2	-130	-149.70	1.34	$\leq \beta \leq$	1.43	-139.85	9.85
	3	-120	-159.00	1.24	$\leq \beta \leq$	1.34	-139.50	19.50
	4	-110	-168.20	1.15	$\leq \beta \leq$	1.24	-139.10	29.10
	5	-100	-177.50	1.05	$\leq \beta \leq$	1.15	-138.75	38.75
	6	-90	-187.20	0.94	$\leq \beta \leq$	1.05	-138.60	48.60
	7	-80	-197.70	0.84	$\leq \beta \leq$	0.94	-138.85	58.85
	8	-70	-209.40	0.74	$\leq \beta \leq$	0.84	-139.70	69.70
	9	-60	-223.00	0.64	$\leq \beta \leq$	0.74	-141.50	81.50
	10	-50	-239.70	0.55	$\leq \beta \leq$	0.64	-144.85	94.85
	11	-40	-261.50	0.45	$\leq \beta \leq$	0.55	-150.75	110.75
	12	-30	-293.00	0.35	$\leq \beta \leq$	0.45	-161.50	131.50
	13	-20	-346.00	0.25	$\leq \beta \leq$	0.35	-183.00	163.00
	14	-10	-468.00	0.00	$\leq \beta \leq$	0.25	-239.00	229.00
stock-	1	-1120	-1120.00	1.43	$\leq \beta \leq$	2.00	-1120.00	0.00
market	2	-1040	-1197.60	1.34	$\leq \beta \leq$	1.43	-1118.80	157.60
	3	-960	-1272.00	1.24	$\leq \beta \leq$	1.34	-1116.00	312.00
	4	-880	-1345.60	1.15	$\leq \beta <$	1.24	-1112.80	465.60
	5	-800	-1420.00	1.05	$\leq \beta \leq$	1.15	-1110.00	620.00
	6	-720	-1497.60	0.94	$\leq \beta \leq$	1.05	-1108.80	777.60
	7	-640	-1581.60	0.84	$\leq \beta \leq$	0.94	-1110.80	941.60
	8	-560	-1675.20	0.74	$\leq \beta \leq$	0.84	-1117.60	1115.20
	9	-480	-1784.00	0.64	$\leq \beta \leq$	0.74	-1132.00	1304.00
	10	-400	-1917.60	0.55	$\leq \beta \leq$	0.64	-1158.80	1517.60
	11	-320	-2092.00	0.45	$\leq \beta \leq$	0.55	-1206.00	1772.00
	12	-240	-2344.00	0.35	$\leq \beta \leq$	0.45	-1292.00	2104.00
	13	-160	-2768.00	0.25	$\leq \beta \leq$	0.35	-1464.00	2608.00
	14	-80	-3744.00	0.00	$\leq \beta \leq$	0.25	-1912.00	3664.00

Note: BAR choice options for the two loss questions. Range indicates the β-parameter range each option implied. Participants only saw columns Outcome 1 and Outcome 2. Information in columns Option (Option Number), "EV" (Expected Value) and "SD" (Standard Deviation) was not provided to participants.

Table A4—BAR Measure Option Values of Mixed Domain

Question	Option	Outcome 1	Outcome 2	low	range	hgh	EV	SD
mixed small	1	10.00	-10	2.96	$\leq \lambda \leq$	3.20	0.00	20.00
casino	2	41.00	-20	2.77	$\leq \lambda \leq$	2.96	10.50	61.00
	3	71.00	-30	2.56	$\leq \lambda \leq$	2.77	20.50	101.00
	4	99.00	-40	2.37	$\leq \lambda \leq$	2.56	29.50	139.00
	5	125.00	-50	2.16	$\leq \lambda \leq$	2.37	37.50	175.00
	6	148.70	-60	1.97	$\leq \lambda \leq$	2.16	44.35	208.70
	7	170.20	-70	1.77	$\leq \lambda \leq$	1.97	50.10	240.20
	8	189.50	-80	1.56	$\leq \lambda \leq$	1.77	54.75	269.50
	9	206.50	-90	1.36	$\leq \lambda \leq$	1.56	58.25	296.50
	10	221.30	-100	1.17	$\leq \lambda \leq$	1.36	60.65	321.30
	11	233.90	-110	0.97	$\leq \lambda \leq$	1.17	61.95	343.90
	12	244.30	-120	0.77	$\leq \lambda \leq$	0.97	62.15	364.30
	13	252.50	-130	0.56	$\leq \lambda \leq$	0.77	61.25	382.50
	14	258.50	-140	0.50	$\leq \lambda \leq$	0.56	59.25	398.50
mixed big	1	80.00	-80	2.96	$\leq \lambda \leq$	3.20	0.00	160.00
stock-market	2	328.00	-160	2.77	$\leq \lambda \leq$	2.96	84.00	488.00
	3	568.00	-240	2.56	$\leq \lambda \leq$	2.77	164.00	808.00
	4	792.00	-320	2.37	$\leq \lambda \leq$	2.56	236.00	1112.00
	5	1000.00	-400	2.16	$\leq \lambda \leq$	2.37	300.00	1400.00
	6	1189.60	-480	1.97	$\leq \lambda \leq$	2.16	354.80	1669.60
	7	1361.60	-560	1.77	$\leq \lambda \leq$	1.97	400.80	1921.60
	8	1516.00	-640	1.56	$\leq \lambda \leq$	1.77	438.00	2156.00
	9	1652.00	-720	1.36	$\leq \lambda \leq$	1.56	466.00	2372.00
	10	1770.40	-800	1.17	$\leq \lambda \leq$	1.36	485.20	2570.40
	11	1871.20	-880	0.97	$\leq \lambda \leq$	1.17	495.60	2751.20
	12	1954.40	-960	0.77	$\leq \lambda \leq$	0.97	497.20	2914.40
	13	2020.00	-1040	0.56	$\leq \lambda \leq$	0.77	490.00	3060.00
	14	2068.00	-1120	0.50	$\leq \lambda \leq$	0.56	474.00	3188.00

Table A5—BAR Scale Mixed

Note: BAR choice options for the two mixed questions. "Range" indicates the λ-parameter range each option implied. Participants only saw columns "Outcome 1" and "Outcome 2". Information in columns "Option" (Option Number), "EV" (Expected Value) and "SD" (Standard Deviation) was not provided to the participants. The λ values are conditional on the assumption that $\alpha = 0.9$ and $\beta = 0.9$.

A2. EG Choice Options

Table A6—Options in the Eckel and Grossman Measure

Choice	Payoff		Expected	Variance	CRRA.Range	
(50/50 Gamble)	Low	High	Return		lower	upper
Gamble 1	28	28	0	0	3.46	$=\infty$
Gamble 2	24	36	6	12	1.16	3.46
Gamble 3	20	44	12	24	0.71	1.16
Gamble 4	16	52	18	36	0.50	0.71
Gamble 5	12	60	24	48	0	0.50
Gamble 6	2	70	34	68	¡ 0	0

Note: Adapted from Charness and Viceisza (2012) *Low* = lower possible payoff when this lottery was chosen, *High* = higher possible payoff when this lottery was chosen. *CRRA Range* = range of risk parameter assuming constant relative risk aversion. The value represents r in a utility function of the form $u(x) = x^{1-r}$ where x is the monetary value of the option.

A3. Mechanical Turk Procedures

On the Mechanical Turk Plattform (mTurk), participants saw the following description, whereupon they could decide to participate in our study:

Dear Madam or Sir,

Thank you for participating in our study. This study is part of an official research project of the University of Basel, Switzerland. With this series of questionnaires, we aim to study participants' risk-taking behavior. You will be asked to fill out several different questionnaires, which will assess your risk taking behavior in various situations. There are no physical or legal risks involved in your participation in this study. The study will take on average 30 minutes. For your participation in the study, you will receive $4.5 (i.e. $9 an hour) plus a bonus of up to 5$ depending on some of the questions. The study is conducted by the Center for Economic Psychology, Department of Psychology, University of Basel, Switzerland. The study consists of two parts. This is the first part.

We comply with the privacy regulations set forth/ specified by the Department of Psychology at the University of Basel. Your data is collected online, stored anonymously and used for scientific purposes only. The results of the study are pooled anonymously and published scientifically without reference to your person.

Participation in the study is voluntary, and you have the opportunity to discontinue your participation at any time. Likewise, you are free at any time to retract your consent for future usage of your data. If you have any questions about this study, you can contact the primary investigator via email, at...

On accepting the assignment, the participant was redirected to the Unipark site, where all questionnaires were programmed in, and the participant was asked to sign a consent form electronically. Next, each participant had to indicate their unique mTurk worker-ID.

All risk questionnaires were presented to the participants in a random order. On page 3 of the survey, we included a question that checked whether a participant had paid attention to the task and carefully read the instructions. The

question was as follows:

Most modern theories of decision making recognize the fact that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables can greatly impact the decision process. To facilitate our research on decision-making, we are interested in knowing certain factors about you, the decision-maker. Specifically, we are interested in whether you take the time to read the directions; if not, then some of our manipulations that rely on changes in the instructions will be ineffective. So to demonstrate that you have read the instructions, please ignore the sports items below and type: "yes, I have" into the text box labeled other. We greatly appreciate your truthfulness and will start with the questions on the next page.

All participants passed this attention check.

After completion of all questionnaires, participants received a number code that they could enter on the mTurk site and complete the assignment. After a maximum of three days, they received the monetary compensation for their participation in the study.

A4. Model-Fitting Procedures with Maximum Likelihood

To evaluate the fit of the models describing the BAR, we used a maximum likelihood estimation (MLE) technique. The likelihood of a model given the data was calculated for each choice of every participant. The likelihood of choosing an option y_i over all I options in a given question was given by the cumulative density function of the normal distribution:

$$P(y_i|y_1, y_2, ..., y_{14}) = \frac{1}{\sqrt{2\pi\sigma}}^{-\frac{(y_i-\mu)^2}{2\sigma^2}}, \tag{A1}$$

where μ denotes the option with the maximum utility, such that

$$\mu = max(u(y_1), u(y_2), ..., u(y_{14})), \tag{A2}$$

and σ denotes the standard deviation of the likelihood function, which we defined as described in the next paragraph.

We used a normal distribution as likelihood function because the choice options in the BAR were designed in a way that every option a participant could take included the same size of range. The standard deviation of the likelihood function (σ) was fixed for all participants and models. The fixation of the standard deviation, or behavioral consistency parameter, is done because of model-fitting stability reasons. Because we estimated the parameters for each participant over only 6 data points, letting the variance of the likelihood be free would affect the other parameter values, especially if they were highly correlated. Thus, to fix the standard deviation, we first ran all models with the standard deviation parameter set as a free parameter estimated from the data and then took the average over all models and participants as a fixed value for the final run of parameter

estimations. In contrast, the mean of the likelihood function μ was dependent on the parameter values. Thus, the maximum likelihood corresponded to the highest probability that the observed data were generated by the model with the particular parameters for which the difference between the utility from the chosen option and the maximum utility across all choice options was the smallest.

For model 1, we follow Wilcox (2011) and normalize all option outcomes to a range of $[0, 1]$ by dividing with the difference between the maximum utility and the minimum utility over all outcomes in each question.

$$x_{norm} = \frac{x - min(x)}{max(x) - min(x)} \tag{A3}$$

This normalization, known as contextual utility, serves to capture lottery-specific heteroskedasticity in the error term Bruhin et al. (2010). One of the main advantages of this normalisation is that it shifts all parameter values into the same *positive* range.

For the other models, we did not need (nor want) the shift into the positive domain but only into a range of [-1,1] and thus normalized the options by the most extreme value (the maximum of the absolute values):

$$x_{norm} = \frac{x}{max(|x|)} \tag{A4}$$

To estimate the parameters, we used a differential evolution algorithm of the "DEoptim" package in R (Mullen et al., 2009) with the following boundaries, based on findings in the previous literature Köbberling and Wakker (2005); Tversky and Kahneman (1992); Nilsson et al. (2011): $\alpha = [0.1, 2], \beta : [0.1, 2], \lambda : [0.1, 5]$. We limited the evolution iterations of the algorithm to 1000 for each participant. We fit the models using the log-likelihoods of each model given the

data and compared the models using the Bayesian Information Criterion (BIC).

Probabilistic Choice Rule. — To compare the model fits, we estimated the likelihood of a model given a set of individual parameter values using a probabilistic choice rule. The choice rule was normally distributed, which provided the likelihood of a single choice given a certain parameter. The normal distribution was drawn over the 14 options of a question in the BAR with the mean μ representing the option with the highest utility for the given set of parameter values. We used the sums of the log-likelihoods for all six questions given a certain model for each participant. The variance of the normally distributed likelihood function represents the behavioral consistency of an individual and can either be estimated from the data individually for each participant or fixed. When estimating a model over only six data points, the estimated parameters can be substantially intercorrelated and thus confound the correct estimation. Thus, we ran the modeling analysis twice. First, with the variance as a free parameter estimated from the data. In a second estimation we fixed the variance for all models and all individuals to the mean of the variance parameter when it was estimated separately. This way we aimed to control for intercorrelations of the parameters. The overall model fits of both estimation techniques did not substantially change the outcomes, which is why we use the estimation with the fixed variance for the rest of the analysis since we assume those estimations are more robust. Detailed results of the modeling results with free estimated variance can be provided at request.

Appendix B: Additional Tables and Figures for Results

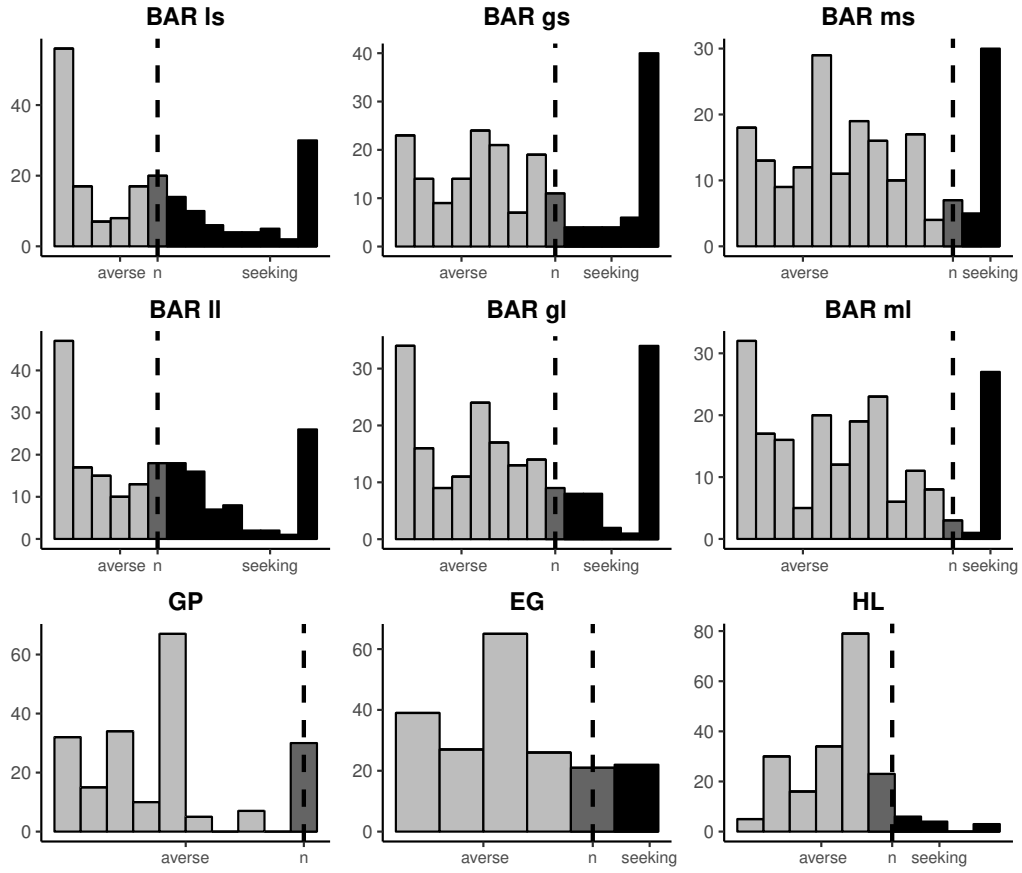

FIGURE B1. DISTRIBUTIONS FOR RAW CHOICES OF LOTTERY EM MEASURES SESSION 1

Note: Distributions of choices in risk-preference measures in the BAR. ls = losses small, ll = losses large, gs = gains small, ls = gains large, ms = mixed small, ml = mixed large. The dashed line indicates the option which represented risk-neutral preference. Thus, light-gray shaded bars represent risk averse subjects, dark-gray shaded bars represent risk-neutral subjects, and black bars represent risk-seeking subjects. Note that each measure offers different number of options in the respective category.

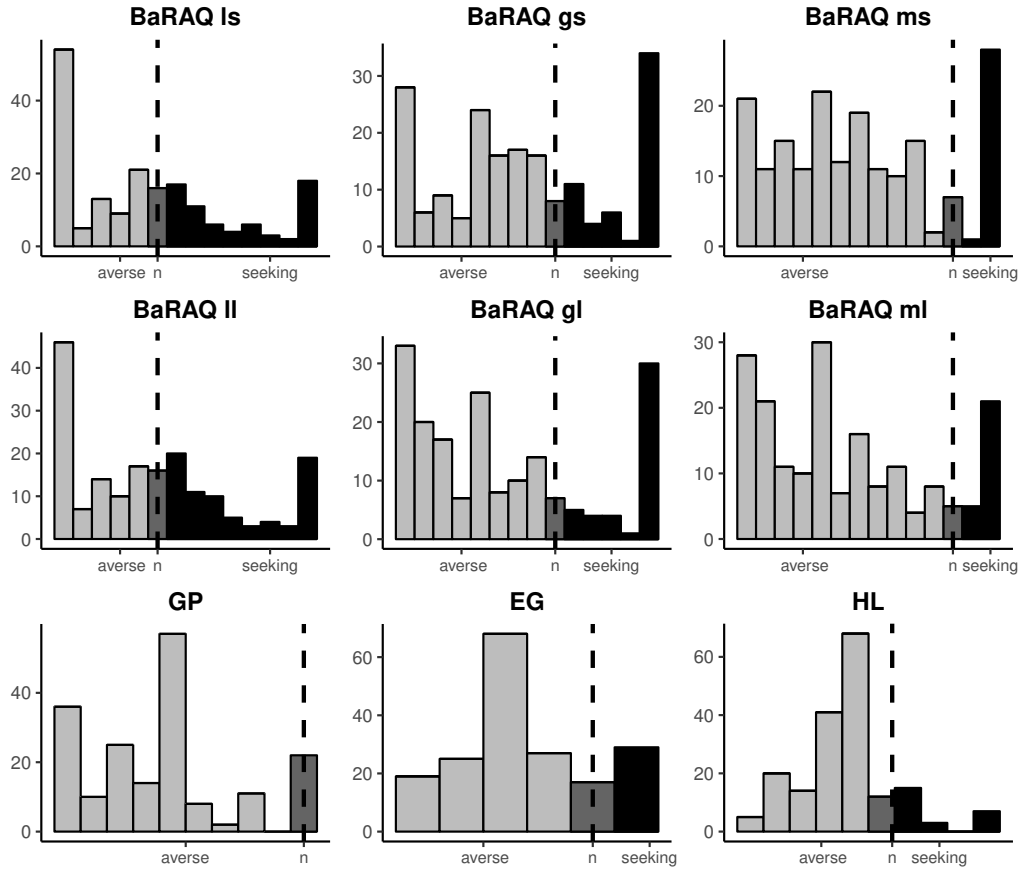

Figure B2. Distributions for raw choices of lottery EM measures Session 2

Note: Distributions of choices in risk-preference measures.BAR = Basel Risk Attitude Scale, ls = losses small, ll = losses large, gs = gains small, ls = gains large, ms = mixed small, ml = mixed large. The dashed line indicates the option which represented risk-neutral preference. Thus, light-gray shaded bars represent risk-averse subjects, dark-gray shaded bars represent risk-neutral subjects, and black bars represent risk-seeking subjects. Note that each measure offers different numbers of options in the respective category.

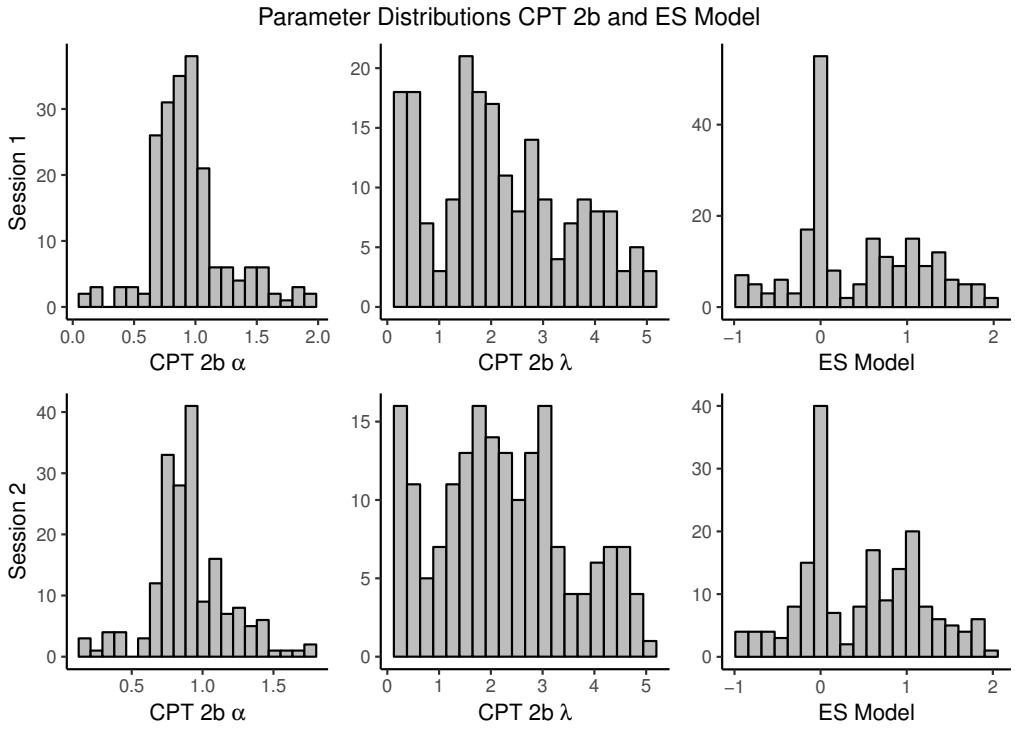

FIGURE B3. DISTRIBUTIONS OF PARAMETER VALUES OF CPT 2B AND EXPECTED SHORTFALL MODEL

Note: Figure shows the distributions of the parameter estimates for the second- and third-best model ES model and CPT 2b.

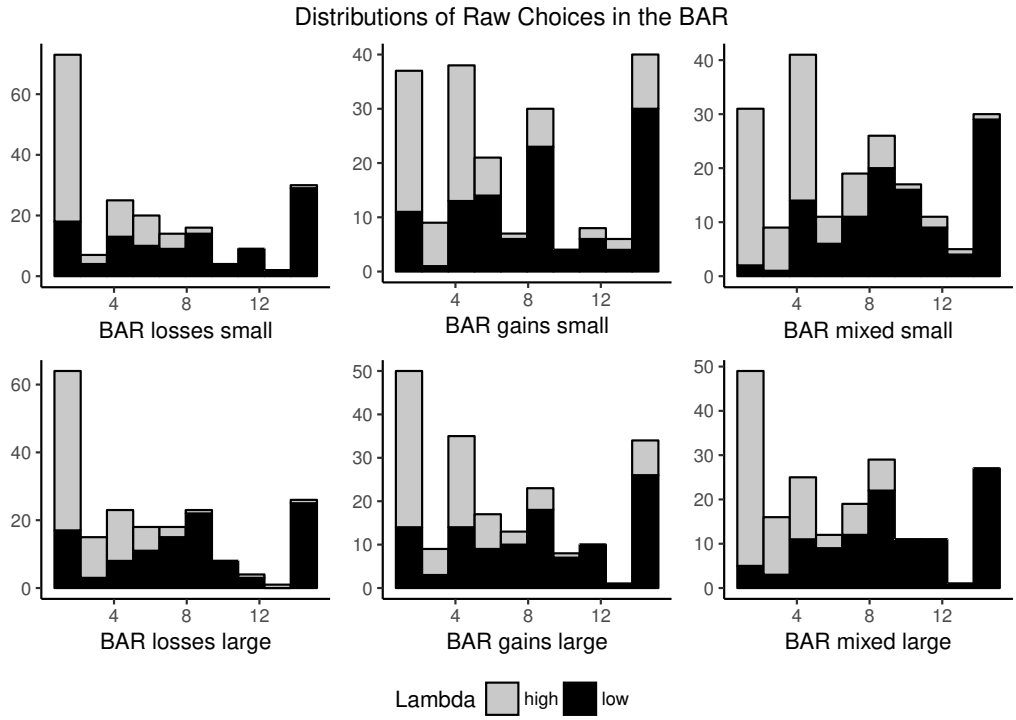

Figure B4. Distributions of Raw Choices in the BAR

Note: Distributions of raw choices in the BAR. The color coding indicates the proportion of participants for which a high value of λ was estimated for the CPT-3 model. Light-gray = λ estimates over 4, black = λ estimates below 4. Note that the participants with high lambdas generally chose risk-averse options (low raw values) for all items.

TABLE A7—ESTIMATED UTILITY PARAMETERS CORRESPONDING TO HL CHOICE OPTIONS

			CRRA Range	
Decision	Option A	Option B	Lower	Upper
1	\$2.00 if the die is 1 \$1.60 if the die is 2 - 10	\$3.85 if the die is 1 \$0.10 if the die is 2 - 10	$-\infty$	-1.71
2	\$2.00 if the die is 1- 2 \$1.60 if the die is 3- 10	\$3.85 if the die is 1- 2 \$0.10 if the die is 3- 10	-1.71	-0.95
3	\$2.00 if the die is 1-3 \$1.60 if the die is4 -10	\$3.85 if the die is 1- 3 \$0.10 if the die is 4 - 10	-0.95	-0.49
4	\$2.00 if the die is 1-4 \$1.60 if the die is 5-10	\$3.85 if the die is 1- 4 \$0.10 if the die is 5 - 10	-0.49	-0.14
5	\$2.00 if the die is 1-5 \$1.60 if the die is 6-10	\$3.85 if the die is 1 - 5 \$0.10 if the die is 6 - 10	-0.14	0.15
6	\$2.00 if the die is 1-6 \$1.60 if the die is 7-10	\$3.85 if the die is 1- 6 \$0.10 if the die is 7 - 10	0.15	0.41
7	\$3.85 if the die is 1 \$0.10 if the die is 2 - 10	\$3.85 if the die is 1- 7 \$0.10 if the die is 8 - 10	0.41	0.86
8	\$2.00 if the die is1-8 \$1.60 if the die is 9-10	\$3.85 if the die is 1- 8 \$0.10 if the die is 9 - 10	0.68	0.97
9	\$2.00 if the die is 1- 9 \$1.60 if the die is 10	\$3.85 if the die is 1- 9 \$0.10 if the die is 10	0.97	1.37
10	\$2.00 if the die is 1_10	\$3.85 if the die is 1_10	1.37	∞

Note: Participants made a choice of Lottery A or B for each of the 10 decisions. The two rightmost columns (CRRA Range) were not presented to the participants. The CRRA range was based on a utility function of the form $x = \frac{x^{1-r}}{1-r}$

Table B1—Correlation Among SOEP Questions

		1	2	3	4	5	6	7	8
1	general	1.000***	0.300***	0.779***	0.580***	0.587***	0.379***	0.323*	0.419
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
2	finance	0.300	1.000	0.387	0.422	0.372	0.470	0.173	0.128
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.058)	(0.210)
3	sport	0.779	0.387	1.000	0.559	0.619	0.413	0.325	0.315
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
4	job	0.580	0.422	0.559	1.000	0.527	0.423	0.187	0.297
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.041)	(0.000)
5	health	0.587	0.372	0.619	0.527	1.000	0.481	0.233	0.199
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.007)	(0.029)
6	social	0.379	0.470	0.413	0.423	0.481	1.000	0.227	0.090
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.009)	(0.409)
7	investment	0.323	0.173	0.325	0.187	0.233	0.227	1.000	0.056
		(0.000)	(0.015)	(0.000)	(0.008)	(0.001)	(0.001)	(0.000)	(0.428)
8	gambling	0.419	0.128	0.315	0.297	0.199	0.090	0.056	1.000
		(0.000)	(0.070)	(0.000)	(0.000)	(0.005)	(0.205)	(0.428)	(0.000)

Note: Values in brackets represent p-vales. The upper estimates are corrected for multiple comparisons. Note that all real-life risk attitudes correlate with general real-life risk attitude.

Table B2—Spearman Correlations Among EM in Session 1

	BAR	HL	GP	EG	SS
BAR					
HL	0.33(<.001)				
GP	0.56(<.001)	0.28(<.001)			
EG	0.39(<.001)	0.36(<.001)	0.33(<.001)		
SS	0.26(<.001)	0.18(0.011)	0.16(0.027)	0.19(0.008)	
IMP	0.14(0.043)	0.07(0.316)	0.14(0.046)	0.15(0.038)	0.30(<.001)

Note: BAR = Basel Risk Measure, HL = Holt and Laury task, GP = Gneezy and Potters task, EG = Eckel and Grossman task, SS = Sensation Seeking Score, IMP = Barratt Impulsiveness Score, SOEP-G = general willingness to take risk in real-life. Values in brackets represent p-values.

Table B3—Spearman Correlations Among EM in Session 2

BAR					
HL	0.37(<.001)				
GP	0.38(<.001)	0.27(<.001)			
EG	0.54(<.001)	0.43(<.001)	0.31(<.001)		
SS	0.32(<.001)	0.13(0.069)	0.16(0.031)	0.23(0.002)	
IMP	0.21(0.005)	0.07(0.315)	0.07(0.313)	0.11(0.133)	0.39(<.001)

Note: BAR = Basel Risk Measure, HL = Holt and Laury task, GP = Gneezy and Potters task, EG = Eckel and Grossman task, SS = Sensation Seeking Score, IMP = Barratt Impulsiveness Score, SOEP-G = general willingness to take risk in real-life. Values in brackets represent p-values.

Table B4—Model Parameter Correlation Between CPT-3 with Optimized vs with Derived Parameters

	CPT-3 α optimized	CPT-3 β optimized	CPT-3 λoptimized
CPT-3 α derived	0.93(<.001)	-0.12 (0.104)	-0.50(<.001)
CPT-3 β derived	-0.23(<.001)	0.85(<.001)	0.72(<.001)
CPT-3 λ derived	-0.45(<.001)	0.28(<.001)	0.85(<.001)

Note: Shows the correlations among the model parameters of the CPT-3 model (MLE estimated parameters) and the READ model (parameters read out from the normed table). Values in parenthesis represent p-values. Correlations are calculated with the Spearman method.

Table B5—Categorization of Risk-Aversion by Measure

	Median r	Risk Averse	Risk Neutral	Risk Seeking
BAR α	0.76	68.50%	21.50%	10.00%
EG	0.60	78.50%	10.50%	11.00%
GP	0.65	85.00%	15.00%	-
HL	0.32	94.00%	3.00%	3.00%

Table B6—Model Comparison with free Behavioral Consistency Parameter

Model	LL1	BIC1	No.bestfit1	LL2	BIC2	No.bestfit2
CPT1	-15.26	36.11	7	-15.24	36.03	3
CPT2a	-12.95	31.51	28	-14.85	35.30	32
CPT2b	-12.96	31.51	15	-13.49	32.53	24
CPT3	-11.90	29.39	73	-11.85	29.25	73
ES Model	-12.17	29.95	77	-11.90	29.36	53

Note: Model comparison when the variance of the likelihood function was not fixed but estimated from the data. Generally, the order of the fits in terms of BIC were the same as in the results with fixed variance. A main difference is that the number of best fits per model was more dispersed over all models. This means that fewer people were described best by the CPT-3 model than with fixed variance.

Short Research Article

Combining General And Specific Measures of Risk Preference Boosts Predictive Power in the Wild

Oliver Schürmann and Renato Frey

University of Basel

Ralph Hertwig

Max Planck Institute for Human Development, Berlin

Jörg Rieskamp

University of Basel

Andreas Pedroni

University of Basel, University of Zurich

Author Note

This research has been made possible through the support of the Swiss National Science Foundation, with grants to the first author (P0BSP1_148884) and the fourth and fifth authors (CRSII1_136227).

Correspondence concerning this article should be addressed to Oliver Schürmann, Center for Economic Psychology, Department of Psychology, Missionsstrasse 62a, 4055 Basel, Switzerland. E-mail: o.schuermann@unibas.ch

Word count Introduction & Discussion: 1689

Abstract

More than one-fifth of people killed on roads worldwide are pedestrians. Pedestrian fatalities occur under various circumstances. Previous research has studied the extent to which individual risk preferences predict risky decisions by pedestrians or other actors in traffic. However, this research has commonly probed risk preference and criterion behavior concurrently and measured risk behavior in a laboratory setting rather than in the wild. We addressed these shortcomings and predicted risky road crossing decisions on an actual street after ten months by harnessing the complementary effects of domain-general and domain-specific risk preference measures. We found that a model integrating domain-general and domain-specific measures had substantially higher predictive accuracy than did single-predictor models. Furthermore, this model also predicted self-reported dangerous driving, thus showing evidence of generalizability. The theoretical implications of these findings are that risk preferences have multiple facets and that integrating them promises better prediction of risky decision-making in the wild.

Keywords: naturalistic risk taking, prediction, risk preference, traffic, street crossing, risk, BART, sensation seeking, impulsivity

Combining General And Specific Measures of Risk Preference Boosts Predictive Power in the Wild

One of the key reasons for measuring individual risk preference is to predict engagement in potentially hazardous behaviors—drug use, unprotected sex, and speeding, to name just a few (Dohmen et al., 2011; Hoffrage, Weber, Hertwig, & Chase, 2003; Hopko et al., 2006). Yet the road from measuring a person's risk preference to predicting potentially perilous decisions in the wild is a bumpy one. Previous studies have shown that risk preference measures have limited predictive accuracy for risky decisions (see Fox & Tannenbaum, 2011; Schonberg, Fox, & Poldrack, 2011). Worryingly, many of these studies probed preference measures and actual risk-taking behavior concurrently—typically on the same day—which may have inflated the estimated predictive accuracy (Dahlen, Martin, Ragan, & Kuhlman, 2005; Rosenbloom, 2006; Vaca et al., 2013): Measuring preference and criterion behavior within such a narrow time window may introduce carry-over effects and demand characteristics (Reis & Judd, 2014). Finally, most studies have measured risky decision-making in a laboratory setting rather than in a real-life setting, meaning that situational influences on behavior are likely neglected. In sum, it remains uncertain to what extent risk preferences reliably predict future risky decisions.

The goal of this study is to address these shortcomings. Specifically, we draw on recent findings showing that risk preference has both domain-general and domain-specific components (Frey, Pedroni, Mata, Rieskamp, & Hertwig, in press; Highhouse, Nye, Zhang, & Rada, 2016) and combine respective measures to predict risky decision-making in an actual street setting after an average of 10 months. Although gauging potentially harmful behaviors in real-life contexts is challenging and effortful and has often been avoided in previous research (Dahlen et al., 2005; Le Bas, Hughes, & Stout, 2015; Weber, Blais, & Betz, 2002), we examine a concrete real-world behavior of high relevance: road-crossing behavior in a naturalistic setting. An intuitive and

natural starting point for predicting a specific risky decision-making in the wild might be to rely on an equivalent domain-specific laboratory measure. We investigate to what extent complementing such a measure (for traffic risk taking) with domain-general measures of risk preference leads to incremental predictive accuracy.

The Psychometric Structure of Risk Preference

A crucial element in predicting actual risky decision-making is how best to conceptualize and measure the construct of risk preference (see Appelt, Milch, Handgraf, & Weber, 2011; Fox & Tannenbaum, 2011; Frey et al., 2017; Schonberg et al., 2011). One view is that risk preference is a unidimensional construct that manifests consistently in behavior across different domains and circumstances (e.g., Jackson, Hourany, & Vidmar, 1972). Another view is that risk preference varies across situations and domains (e.g., Weber et al., 2002; Figner & Weber, 2011). Finally, a third and more recent view is that risk preference—like intelligence (Spearman, 1904)—has both general and specific components (see Frey et al., 2017; Highhouse et al., 2016). To the extent that actual decisions are shaped by a person's risk preference, combining measures with differing degrees of specificity may thus increase the predictive accuracy of risk-preference assessment. In other words, a combination of domain-specific and relatively general measures of risk preference may be a promising solution to the problem of poor predictive accuracy observed in previous studies.

Measures of Risk Preference and Their Aggregation

Among the available measures of risk preference, some were specifically developed to elicit domain-specific preferences (Weber et al., 2002), whereas others were designed without this goal in mind and have been found to generalize relatively well across domains. We refer to the former as "domain-specific" and to the latter as "domain-general" measures. Domain-specific measures of risk preference have been observed to correlate with specific behaviors, such as

unsafe driving (Hergovich, Arendasy, Sommer, & Bognar, 2007) or risk taking in the domains of recreation or health (Hanoch, Johnson, & Wilke, 2006). In contrast, domain-general measures of risk preference have been found to correlate with a variety of risky behaviors. For example, scores in the Balloon Analogue Risk Task (BART; Lejuez et al., 2002) and the Devil's Task (aka Slovic's Risk Task; Slovic, 1966) correlate with risk taking in behaviors as diverse as road crossing (Hoffrage et al., 2003; Vaca et al., 2013), drug abuse (Skeel, Pilarski, Pytlak, & Neudecker, 2008), substance abuse (Hopko et al., 2006), and gambling (Mishra, Lalumière, & Williams, 2010). Finally, some measures may be indicative of both domain-specific and domain-general behaviors depending on their level of analysis. For example, the Sensation Seeking Questionnaire (Zuckerman, 2007) yields a total score that has been found to correlate with risk taking in road traffic, financial risk taking, engagement in thrill-seeking sports, unsafe sex, and drug abuse (see, e.g., Lauriola et al., 2013, for an overview). Yet the questionnaire can also be decomposed into five subscales that have been found to correlate with more specific types of risk-taking behavior (see Jonah, 1997; Zuckerman, 1994).

In sum, the available measures differ substantially in their degree of specificity. To the extent that the construct of risk preference appears to involve a general component as well as (domain-) specific components, it seems reasonable to expect that complementing a highly domain-specific measure of traffic risk taking with relatively general measures will increase predictive accuracy for actual risky decisions. Furthermore, aggregating multiple measures is known to reduce measurement error. Specifically, combination of predictors that are not perfectly correlated has been shown to boost predictive accuracy substantially (Armstrong, 2001: Ernst et al., 2007; Wolfers & Zitzewitz, 2004).

Overview of the Present Study

Our study aimed at testing the hypothesis that combining domain-general and domain-

specific measures of risk preference will boost predictive accuracy with respect to individual risk taking in road-crossing behavior. In a first (Session 1) of two experimental sessions (see Table 1 for an overview), participants completed 14 behavioral risk preference measures on the computer. Of these measures, we selected four that have previously been shown to correlate with risk taking in traffic as predictor variables. These four measures differ substantially in their degree of specificity, One measure targets behavior specific to the target domain, the Vienna Risk-Taking Test: Traffic (VRTT; Hergovich et al., 2007), whereas the three other measures target relatively domain-general risk-taking tendencies: the Sensation-Seeking Scale V (SS; Zuckerman, 2007), the Barratt Impulsiveness Scale (IMP; Barratt, 1965), and the Balloon Analogue Risk Task (BART; Lejuez et al., 2002). See Supplemental Material A for details of the selection and categorization of the measures.

On average 10 months after the assessment of the risk preferences, participants were invited to a follow-up session (Session 2), in which we measured risky street crossing behavior on an actual street. In addition to using models that combine multiple measures to predict risk taking in street crossing, we sought to test whether our predictions generalized to other dangerous traffic behavior (i.e., driving behavior) that may occur due to increased risk taking.

Table 1

Overview of the Assessments Implemented in the Experimental Sessions

Session 1	Session 2 (10 months later)
Vienna Risk-Taking Test: Traffic (VRTT)	Street crossing task
Sensation-Seeking Scale V (SS)	Self-reported dangerous driving behavior
Barratt Impulsiveness Scale (IMP)	
Balloon Analogue Risk Task (BART)	

Methods

Participants and Procedure

Of a total sample of 1,507 healthy participants who completed a battery of 14 computerized risk preference measures (Session 1; for details, see Frey et al., 2017), 50 participants (68% females, mean age = 25.3 years) were randomly selected and invited to attend a second session (Session 2) that took place on average 10 months later (M = 10.4 months, SD = 2.7 months, min = 6 months, max = 16.3 months). A power analysis using expected effect sizes of r~.35, conducted before session 2, revealed a suggested sample size of around n=60. However, due to the naturalistic nature of the street crossing task, a very low response rate (<40%) for session 2 and the requirement to have a driver license for at least two years, it was not feasible to continue data collection after we collected the data from 50 participants. In Session 2, participants provided informed consent and received written and oral instructions and safety guidance concerning the street-crossing task they were about to conduct. They were then taken to the side of an actual street outside the laboratory (see Figure 1) and performed the street-crossing task, which lasted approximately 20 minutes. Subsequently, they completed the Manchester Driver Behavior Questionnaire on a computer in the laboratory (DBQ; Reason, Manstead, Stradling, Baxter, & Campbell, 1990). Finally, participants were paid 20 Swiss francs (USD: 20.05) for their participation of approximately 60 min. Session 1 one was approved by the local ethical review board (Ethikkommission beider Basel, EKBB; EK:342/12), Session 2 was approved by the institutional review board of the Department of Psychology at the University of Basel.

Predictor Variables: Risk Preference Measures

Of the 14 risk preference measures implemented in Session 1, we selected those that have previously been shown to be related to risk taking in traffic, which resulted in four measures. For

an elaboration of the selection and categorization of the measures see supplemental material A. These four measures were used to predict risk taking in the street-crossing task in Session 2.

Vienna Risk-Taking Test: Traffic (VRTT). Participants' risk preference for *domain-specific risk* (i.e., in traffic situations) was assessed using the VRTT (Hergovich et al., 2007). In this test, participants indicate their propensity to take risk in 24 videos of critical road traffic driving situations. In each video, the participant decides whether or not to execute a potentially dangerous maneuver, such as turning left at a crossroads or overtaking a car with another vehicle approaching in the other lane. The participant first watches the whole scene and then indicates at which point the maneuver becomes too dangerous to execute in a second presentation of the video. The later the participant presses the button, the higher the risk of causing a hypothetical accident. The latency from the button press to the end of the video indicates propensity to take risk. A total score is calculated by averaging the latencies of all 24 videos, with larger scores indicating higher risk aversion.

Sensation Seeking (SS). The propensity for sensation seeking, a personality construct that is generalizable over many domains, was assessed using the 40-item Sensation-Seeking Scale V (Zuckerman, 2007). For each item, the participant chooses between descriptions of two actions, one of which involves more sensation seeking (" I like to have new and exciting experiences and sensations even if they are a little frightening, unconventional, or illegal.") than the other (e.g. "I am not interested in experiences for its own sake"). The actions take place in different contexts to generalize over multiple domains. The SS score was quantified by the number of choices representing sensation-seeking actions.

Barratt Impulsiveness Scale (IMP). Impulsivity, the tendency to act without forethought, has been associated with domain-general risk preference (Appelt et al., 2011; Lauriola et al., 2013). Impulsivity was measured using the Barratt Impulsiveness Scale (BIS-11a;

Barratt, 1965), a widely used 30-item scale (e.g. “I do things without thinking”, or “I am restless in the theater or in lectures”). Participants rated the frequency of engaging in the described behavior on a 4-point Likert scale ranging from 1 (“rarely/never”) to 4 (“almost/always”). We used the sum of these ratings as the IMP score.

Balloon Analogue Risk Task (BART). The BART (Lejuez et al., 2002) is a behavioral measure of risk preference that has been shown to generalize to many domains of risk taking (see Lauriola et al., 2013). In this computerized task, the participant pumps up 30 balloons, one after the other, by clicking a button. For every pump, the participant receives a point on a temporary account. If the participant decides to stop before a balloon explodes, the points accrued are transferred to a permanent account. If the balloon explodes, all points accrued in that trial are lost. The participant is not informed about the probabilistic structure that governs the explosion of the balloon but is told it will explode at a random point that differs in every trial. In our version, the probability of the balloon exploding was 1/128 at the first pump, 1/127 at the second pump, 1/126 at the third pump and so on until the 128th pump, where it exploded with certainty. Following these underlying probabilities, the 30 explosion points were defined in advance for the 30 trials.

The risk preference measure typically used in the BART is the average number of pumps in all trials in which the balloon did not explode. However, because the task incorporates a learning process that arguably is not part of a person’s risk preference, we instead used a computational cognitive model to decompose behavior in the task, namely, the Bayesian Sequential Risk-Taking Model (BSR) developed by Wallsten, Pleskac, and Lejuez (2005). The model enables the cognitive processes underlying risky decisions in the BART to be factored out (Rolison, Hanoch, & Wood, 2012). We fitted four versions of the BSR model to the data for each participant individually. Details of the models and fitting procedures are provided in Supplemental Material

B. We also refer readers to Wallsten et al. (2005) and Pleskac (2008) for detailed explanations of the BSR model. From the best-fitting model (according to the Bayesian Information Criterion), we extracted the risk sensitivity parameter as a measure of risk preference and used it as a predictor for real-life risk taking.

Dependent Variables: Dangerous Road Behavior (Session 2)

Street-Crossing Task (SCT). We measured risk-taking decisions in a naturalistic street-crossing situation. Our street-crossing task is similar to that implemented by Hoffrage et al. (2003). The participant stands on the side of a street (see Figure 1) on which there is a single oncoming vehicle (car or tram). As the vehicle approaches, the participant indicates the moment that he or she would *last* feel safe to cross the street—as in the VRTT described above. For safety reasons, participants never actually crossed the street but indicated their decision by taking a step forward. The experimenter used a stopwatch to time the latency from the initiation of crossing (i.e., the participant's step forward) to the arrival of the car or tram (i.e., when its front bumper crossed the participant's hypothetical crossing path). Thus, the shorter the latency time, the more dangerous the street-crossing behavior; larger scores indicate higher risk aversion. To factor out different vehicle speeds and to increase reliability, we ran the task for ten cars and four trams. Additionally, we recorded two possible confounding variables: (i) road conditions (wet or dry) and (ii) weather conditions. The averaged latencies for trams and cars differed substantially within participants (latency times for trams were higher than for cars) but were highly correlated across participants ($r = .88, p < .001$). We therefore normalized (i.e., z-transformed) and averaged both values to create a single SCT measure. For a second, more intuitive measure of street crossing risk, we categorized participants as risk-takers if they caused a hypothetical accident or a close call on two or more occasions during the car scenario. For this, we calculated the time it

takes a person to cross the street using the average speed of a person (3.5 m/s) and the distance to cross the street (~6m) plus one second for close calls. Participants that chose leeway times shorter than this time (5s) at least twice during the experiment were categorized as high risk-takers.

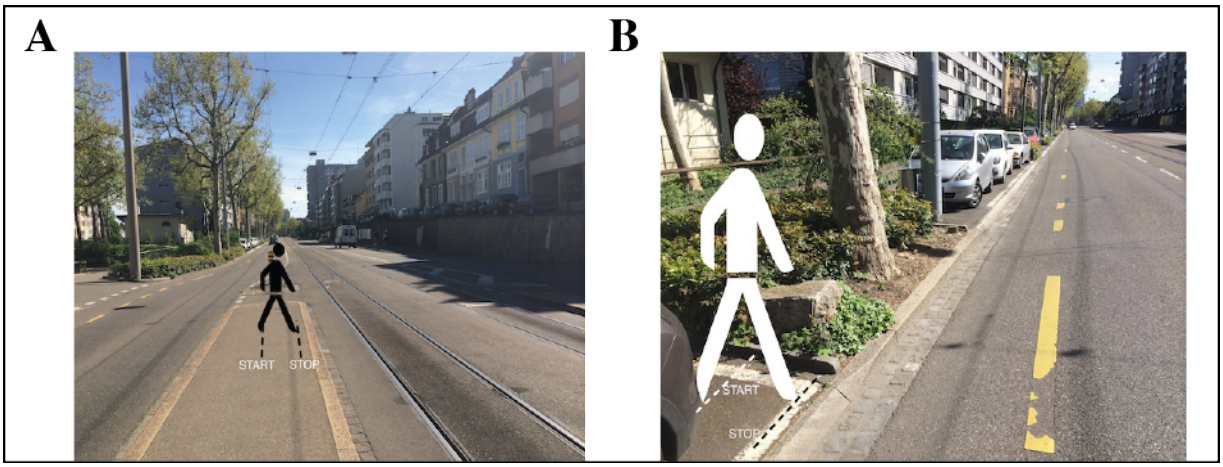

Figure 1. The two SCT situations: **A**: tram; **B**: car. Mannequins represent the participant's standing position.

Self-reported dangerous driving. We assessed self-reported dangerous driving to validate whether the prediction from the Session 1 measurements to actual street-crossing decisions in Session 2 generalized to other dangerous behaviors in traffic. For this purpose, participants completed the Manchester Driver Behavior Questionnaire (DBQ; Reason et al., 1990), consisting of 17 questions about their driving behavior in the past year. The questionnaire comprises three scales reflecting differing intentions: *errors*, *slips/lapses*, and *violations*. *Errors* are defined as "the failure of planned actions to achieve their intended consequences," *slips/lapses* concern failures of attention and memory, and *violations* are "deliberate deviations from those practices believed necessary to maintain the safe operation of a potentially hazardous system" (Reason et al., 1990). Whereas errors and slips/lapses reflect unintended behavior (e.g.,

"braking too quickly on a slippery road"), violations reflect deliberate and typically risk-taking behavior (e.g., "overtaking a slow driver on the inside" or "crossing a junction knowing that the traffic lights have already turned against you"). Thus, only the violation score reflects risk taking in the actual sense; the other two scores should not be dependent on an individual's risk-taking preference.

Statistical Analyses

To test whether predictive accuracy increased meaningfully when the domain-specific measure of risk preference (VRTT) was complemented by domain-general predictor variables, we estimated and compared regression models incorporating all main effects and interactions. To compare these models, we used Bayes factors (BF), a Bayesian alternative to classical hypothesis testing (Rouder, Morey, Speckman, & Province, 2012; Vandekerckhove, Tuerlinckx, & Lee, 2011). The BF is a model comparison statistic that quantifies the support for one model over another. It accounts for goodness-of-fit and model complexity simultaneously, while providing a clearly interpretable measure of the evidence for one model over another. Moreover, the BF can quantify the evidence for the null model (i.e., that there is no effect). For instance, a BF of 1 indicates equal evidence for both models, whereas a BF of 10 (0.1) implies that the data is ten times more likely under model 1 than under model 2 (under model 2 than under model 1). The BFs for the multiple regressions were computed using version 0.9.9 of the BayesFactor R package with the default prior settings (Rouder et al., 2012). We typically compared the models under examination with the null model (i.e., the model incorporating only the intercept). We refer to these null model comparisons as $BF_{0.}$

Results

We first examined to what extent the four risk preference measures assessed in Session 1 correlated with each other. Table 2 reports the coefficients for each correlation and the

corresponding comparisons with the null model (BF_0). Except for the correlation of SS and IMP, for which we found positive evidence ($r = 0.34$, $BF_0 = 3.30$), there was no evidence for correlations between the risk preference measures (BF_0: 0.29 to 0.44). In stark contrast, all risk preference measures were at least moderately correlated with real-world risk taking in the SCT in Session 2 (Table 2, lower panel). As expected, the domain-specific preference measure (VRTT) showed the highest correlation with street-crossing risk, followed by SS, IMP, and the BART. In further analyses, we confirmed that no potential confounding variables (i.e., weather conditions, interval between the two sessions, gender, age, and average walking time in traffic per week) decisively influenced risk taking in the SCT (for details, see Supplemental Material C).

Table 2

Correlations, Means and Standard Deviations of Risk Measures

	Descriptives		Pearson Correlations							
Measure	Mean	SD	VRTT		SS		IMP		BART	
			r	BF_0	*r*	BF_0	*r*	BF_0	*r*	BF_0
			Risk-Preference Measures							
VRTT	7.8s	1.7s	–	–	.08	*0.32*	.14	*0.42*	.15	*0.44*
SS	62.1	7			–	–	**.34**	*3.30*	.04	*0.29*
IMP	63.9	10.4					–	–	.07	*0.30*
BART	35.7	12.03							–	–
			Street-Crossing Risk							
SCT (for cars/trams)	6.9s/9.5s	1.7s/3.3s	**.45**	*27.49*	**.34**	*3.85*	**.30**	*1.77*	**.28**	*1.45*

Note. BF_0 are model calculations against the intercept-only model, the null model typically used to calculate BF statistics

We next examined to what extent predictive accuracy increased when the domain-specific measure (VRTT) was complemented by domain-general measures (SS, IMP, BART). To this end, we estimated regression models for all combinations of the general measures (SS, IMP, BART) with the VRTT (see Figure 2A). For all composite models, the data showed at least strong evidence against the null model. Additionally, no composite model fitted substantially worse than the model with domain-specific risk preference alone (VRTT model)—remember that the BF already punishes for model complexity (see Figure 2A).[1] Apart from the combination of VRTT and BART, which matched the performance of the VRTT-only model, all other models showed at least some evidence of being a better fit than the VRTT-only model. Two models showed the best performance in terms of prediction accuracy and BF_0 (adjusting for model parsimony): one model complemented domain-specific preference (VRTT) by SS and the BART; the other by SS only. The VRTT+SS+BART model performed slightly better ($R^2 = 0.39$, $r = 0.62$, $BF_0 = 563.9$) than the VRTT+SS model ($R^2 = 0.35$, $r = 0.58$, $BF_0 = 560.8$; see Figure 2C). However, comparing both models by means of the *BF* revealed that neither of the two best-fitting models had a significant advantage ($BF = 1.01$). Hence, we selected the more parsimonious, two-variable model (VRTT+SS) to further validate the generalizability of our predictions to other dangerous road behavior. Most importantly, Panels B and C of Figure 2 show scatterplots of the VRTT and VRTT+SS model predicting risk taking in the SCT. Direct comparison of the two models revealed strong evidence for the composite model outperforming the VRTT-only model ($\Delta R^2 = .15$, $BF = 20.51$).

As a second analysis, we tested how well the VRTT+SS model identified those who took the risk of causing an accident or a close call in the car scenario (high risk-takers). As mentioned,

[1] Supplemental Material D shows detailed results for all composite model comparisons, including all linear combinations of main effects and interaction terms.

we classified participants as high risk takers if they chose leeway times shorter than the average time to cross the street plus one second on two or more occasions. 15 participants were classified as risk-takers. A median split of the predictions of the VRTT+SS model identified 12 of the 15 risk-takers to be in the high risk-group (80%).

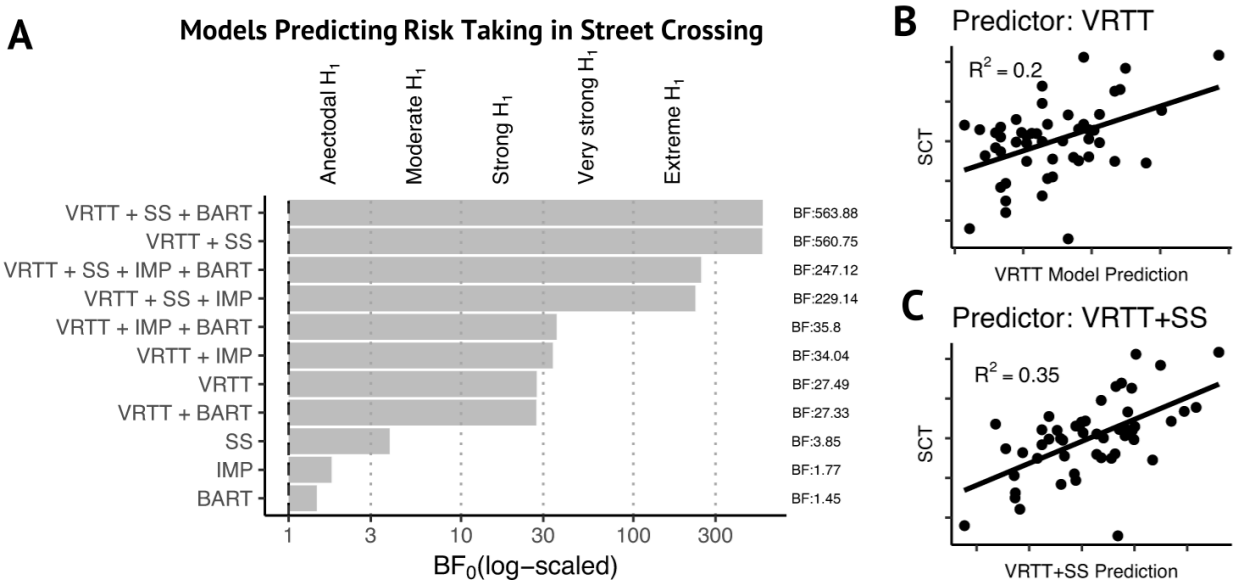

Figure 2. **A**: Comparing the ability of models with single risk preference measures and composites of domain-general and domain-specific measures to predict risk taking in the SCT. All comparisons are in terms of BF_0 (intercept-only model). **B:** Scatterplot of the VRTT model's predictions of risk taking in the SCT ($R^2 = 0.20, r = 0.45, BF_0 = 27.49$) **C:** Scatterplot of the VRTT+SS model's predictions of risk taking in the SCT ($R^2 = 0.35$ $r = 0.62, BF_0 = 563.9$).

In a final step, we tested how well the best-performing model, that is, the VRTT+SS model, generalized to dangerous road behavior other than crossing a street by correlating the predictions of the estimated model with DBQ scores. This procedure represents a quasi-independent validation of the model, because we did not fit the model to predict the DBQ scores,

but instead used predictions for street crossing. Participants with higher predicted risk-taking values had more reported *violations* than individuals with lower predicted risk-taking values (r = .27, BF_0 = 2.64). In contrast, predicted risk-taking values were not related to dangerous driving due to unintended *errors* (r = .15, BF_o = 0.44) or *slips/lapses* (r = .16, BF_0 = 0.49).

Discussion

Our goal was to address shortcomings in past research exploring the predictive accuracy of risk preference measures for potentially risky decisions in the wild. Specifically, we combined one domain-specific (the VRTT) and three domain-general risk preference measures (the BART, IMP, and SS) to predict future risky decisions in a naturalistic street-crossing task. The domain-general risk preference measures were not correlated with the domain-specific measure, but all measures were substantially correlated with the target criterion. Consequently, the best composite models predicted risky decisions substantially better than any single predictor model. This finding is in line with findings from other fields, where combining independent predictors substantially raises predictive accuracy (Armstrong et al., 2001). The best model was the one combining the traffic component of the VRTT with the SS score. It predicted a substantial amount of the variance in street-crossing behavior (R^2 = .35) and was also able to predict self-reported dangerous driving behavior, thus showing evidence of generalizability. To our knowledge, our investigation is the first to harness the power of composite models to predict naturalistic future behavior. The 10-month gap between assessment of risk preferences and risk behavior is likely to have minimized carry-over effects and demand characteristics (Reis & Judd, 2014) and, importantly, it speaks for the temporal reliability of the predictions.

Our findings show that all four of the risk preference measures applied moderately predicted future risk taking in a naturalistic setting. These findings are consistent with other studies, which have found correlations between risk preferences and risk taking in a laboratory

setting (BART: Vaca et al., 2013; SS: Dahlen et al., 2005; Freeman & Rakotonirainy 2015; Jonah, 1997; Rosenbloom, 2006; IMP: Cheng, Ng, & Lee, 2012; Dahlen et al., 2005; VRTT: Hergovich, 1997; Vingilis et al., 2015). To our knowledge, no previous studies of risk-taking propensity in the wild have combined domain-specific and domain-general measures of risk preference in order to improve prediction accuracy.

Although some composite models fitted the data better than single-predictor models, not all combinations of predictors yielded substantially better fits to the data. In contrast to SS, combining either IMP or BART with the domain-specific predictor (VRTT) did not result in an increase in predictive accuracy—potentially because IMP and BART were only moderately correlated with street-crossing risk in the first place. The fact that the BART did not increase the predictive accuracy can be partially explained by recent findings about the inconsistency of behavioral risk preference measures (Pedroni, Frey, Bruhin, Dutilh, Hertwig, & Rieskamp, in press).

Our results have two implications, one practical and one theoretical. On a practical level, the findings can help to spot pedestrians (and car drivers) who are more likely to engage in risky traffic behaviors and will therefore be more prone to accidents. On a theoretical level, they support the recently emerging view that risk preference is neither strictly domain specific nor domain general, but may comprise multiple and measure-dependent facets of risk preference (see also Frey et al., 2017; Highhouse et al., 2016; Schmitz et al., 2016). From this perspective, individuals are characterized by both domain-general and domain-specific risk preferences that may converge or diverge in specific domains. The predictive accuracy of risk measures may thus best be fostered by smartly combining proximal and more distal measures of the risk behavior in question. This is an important avenue for future study.

Author Contributions

O.Schürmann, A. Pedroni, J. Rieskamp developed the study concept and designed the study. Testing and data collection was performed by A. Pedroni and R. Frey (session 1) and O. Schürmann (session 2). O. Schürmann and A. Pedroni analyzed and interpreted the data. O. Schürmann, A.Pedroni, R. Frey, R. Hertwig and J. Rieskamp wrote the paper. All authors approved the final version of the manuscript for submission.

Acknowledgments

We thank Susannah Goss for editing the manuscript and providing helpful comments. We also thank Monica Gschwind for managing the data collection in session 1.

The authors declare that they have no conflicts of interest with respect to their authorship or the publication of this article.

Note:

Data on which the manuscripts' analyses were based available to reviewers: https://osf.io/tcy4b/?view_only=4789c5417f2d4cb8ab6775987c8af1fd

References

Appelt, K. C., Milch, K. F., Handgraaf, M. J., & Weber, E. U. (2011). The Decision Making Individual Differences Inventory and guidelines for the study of individual differences in judgment and decision-making research. *Judgment and Decision Making*, *6*(3), 252. Retrieved from: http://journal.sjdm.org/11/11218/jdm11218.html

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners* (Vol. 30). Springer Science & Business Media. http://dx.doi.org/10.1007/978-0-306-47630-3

Barratt, E. . (1965). Factor analysis of some psychometric measures of impulsiveness and anxiety. *Psychological Reports*, *16*(2), 547–554. http://dx.doi.org/10.2466/pr0.1965.16.2.547

Cheng, A. S. K., Ng, T. C. K., & Lee, H. C. (2012). Impulsive personality and risk-taking behavior in motorcycle traffic offenders: A matched controlled study. *Personality and Individual Differences*, *53*(5), 597–602. https://doi.org/10.1016/j.paid.2012.05.007

Dahlen, E. R., Martin, R. C., Ragan, K., & Kuhlman, M. M. (2005). Driving anger, sensation seeking, impulsiveness, and boredom proneness in the prediction of unsafe driving. *Accident Analysis & Prevention*, *37*(2), 341–348. https://doi.org/10.1016/j.aap.2004.10.006

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, *9*(3), 522–550. http://dx.doi.org/10.1111/j.1542-4774.2011.01015.x

Ernst, B., Oakleaf, B., Ahlstrom, M. L., Lange, M., Moehrlen, C., Lange, B., … Rohrig, K. (2007). Predicting the Wind. *IEEE Power and Energy Magazine*, *5*(6), 78–89. https://doi.org/10.1109/MPE.2007.906306

Figner, B., & Weber, E. U. (2011). Who Takes Risks When and Why?: Determinants of Risk Taking. *Current Directions in Psychological Science*, *20*(4), 211–216. https://doi.org/10.1177/0963721411415790

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

Fox, C. R., & Tannenbaum, D. (2011). The Elusive Search for Stable Risk Preferences. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00298

Freeman, J., & Rakotonirainy, A. (2015). Mistakes or deliberate violations? A study into the origins of rule breaking at pedestrian train crossings. *Accident Analysis & Prevention*, *77*, 45–50. https://doi.org/10.1016/j.aap.2015.01.015

Hanoch, Y., Johnson, J. G., & Wilke, A. (2006). Domain Specificity in Experimental Measures and Participant Recruitment: An Application to Risk-Taking Behavior. *Psychological Science*, *17*(4), 300–304. https://doi.org/10.1111/j.1467-9280.2006.01702.x

Hergovich, A., Arendasy, M. E., Sommer, M., & Bognar, B. (2007). The Vienna Risk-Taking Test - Traffic: A New Measure of Road Traffic Risk-Taking. *Journal of Individual Differences*, *28*(4), 198–204. https://doi.org/10.1027/1614-0001.28.4.198

Highhouse, S., Nye, C. D., Zhang, D. C., & Rada, T. B. (2016). Structure of the Dospert: Is There Evidence for a General Risk Factor? *Journal of Behavioral Decision Making*, n/a-n/a. https://doi.org/10.1002/bdm.1953

Hoffrage, U., Weber, A., Hertwig, R., & Chase, V. M. (2003). How to Keep Children Safe in Traffic: Find the Daredevils Early. *Journal of Experimental Psychology: Applied*, *9*(4), 249–260. https://doi.org/10.1037/1076-898X.9.4.249

Hopko, D. R., Lejuez, C. W., Daughters, S. B., Aklin, W. M., Osborne, A., Simmons, B. L., & Strong, D. R. (2006). Construct Validity of the Balloon Analogue Risk Task (BART): Relationship with

MDMA Use by Inner-City Drug Users in Residential Treatment. *Journal of Psychopathology and Behavioral Assessment*, *28*(2), 95–101. https://doi.org/10.1007/s10862-006-7487-5

Jackson, D. N., Hourany, L., & Vidmar, N. J. (1972). A four-dimensional interpretation of risk taking. *Journal of Personality*, *40*(3), 483–501. http://dx.doi.org/10.1111/j.1467-6494.1972.tb00075.x

Jonah, B. A. (1997). Sensation seeking and risky driving: a review and synthesis of the literature. *Accident Analysis & Prevention*, *29*(5), 651–665. http://dx.doi.org/10.1016/S0001-4575(97)00017-1

Lauriola, M., Panno, A., Levin, I. P., & Lejuez, C. W. (2013). Individual Differences in Risky Decision Making: A Meta-analysis of Sensation Seeking and Impulsivity with the Balloon Analogue Risk Task: Personality and Risky Decision Making. *Journal of Behavioral Decision Making*, *27*(1), 20–36. https://doi.org/10.1002/bdm.1784

Le Bas, G. A., Hughes, M. A., & Stout, J. C. (2015). Utility of self-report and performance-based measures of risk for predicting driving behavior in young people. *Personality and Individual Differences*, *86*, 184–188. https://doi.org/10.1016/j.paid.2015.05.034

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., … Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84. https://doi.org/10.1037//1076-898X.8.2.75

Mishra, S., Lalumière, M. L., & Williams, R. J. (2010). Gambling as a form of risk-taking: Individual differences in personality, risk-accepting attitudes, and behavioral preferences for risk. *Personality and Individual Differences*, *49*(6), 616–621. https://doi.org/10.1016/j.paid.2010.05.032

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (in press). The risk elicitation puzzle. *Nature Human Behavior*.

Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 167–185. https://doi.org/10.1037/0278-7393.34.1.167

Ralston, H. J. (1976). Energetics of Human Walking. In R. M. Herman, S. Grillner, P. S. G. Stein, & D. G. Stuart (Eds.), *Neural Control of Locomotion* (pp. 77–98). Springer US. https://doi.org/10.1007/978-1-4757-0964-3_5

Reason, J., Manstead, A., Stradling, S., Baxter, J., & Campbell, K. (1990). Errors and violations on the roads: a real distinction? *Ergonomics*, 33(10–11), 1315–1332. http://dx.doi.org/10.1080/00140139008925335

Reis, H. T., & Judd, C. M. (2013). *Handbook of research methods in social and personality psychology*. Cambridge University Press . http://dx.doi.org/10.1017/CBO9780511996481

Rolison, J. J., Hanoch, Y., & Wood, S. (2012). Risky decision making in younger and older adults: The role of learning. *Psychology and Aging,* 27(1), 129–140. https://doi.org/10.1037/a0024689

Rosenbloom, T. (2006). Sensation seeking and pedestrian crossing compliance. *Social Behavior and Personality: An International Journal*, *34*(2), 113–122. http://dx.doi.org/10.2224/sbp.2006.34.2.113

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Schmitz, F., Manske, K., Preckel, F., & Wilhelm, O. (2016). The Multiple Faces of Risk-Taking. *European Journal of Psychological Assessment*, *32*(1), 17–38. https://doi.org/10.1027/1015-5759/a000335

Schonberg, T., Fox, C. R., & Poldrack, R. A. (2011). Mind the gap: bridging economic and naturalistic risk-taking with cognitive neuroscience. *Trends in Cognitive Sciences*, *15*(1), 11–19. https://doi.org/10.1016/j.tics.2010.10.002

Skeel, R. L., Pilarski, C., Pytlak, K., & Neudecker, J. (2008). Personality and performance-based measures in the prediction of alcohol use. *Psychology of Addictive Behaviors*, *22*(3), 402–409. https://doi.org/10.1037/0893-164X.22.3.402

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Vaca, F. E., Walthall, J. M., Ryan, S., Moriarty-Daley, A., Riera, A., Crowley, M. J., & Mayes, L. C. (2013). Adolescent Balloon Analog Risk Task and behaviors that influence risk of motor vehicle crash injury. *Annals of Advances in Automotive Medicine*, *57*, 77.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62. https://doi.org/10.1037/a0021765

Vingilis, E., Roseborough, J. E. W., Wiesenthal, D. L., Vingilis-Jaremko, L., Nuzzo, V., Fischer, P., & Mann, R. E. (2015). Experimental examination of the effects of televised motor vehicle commercials on risk-positive attitudes, emotions and risky driving inclinations. *Accident Analysis & Prevention*, *75*, 86–92. https://doi.org/10.1016/j.aap.2014.11.008

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling Behavior in a Clinically Diagnostic Sequential Risk-Taking Task. *Psychological Review*, *112*(4), 862–880. https://doi.org/10.1037/0033-295X.112.4.862

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*(4), 263–290. https://doi.org/10.1002/bdm.414

Wolfers, J., & Zitzewitz, E. (2004). Prediction Markets. *The Journal of Economic Perspectives*, *18*(2), 107–126. https://doi.org/10.1257/0895330041371321

Zuckerman, M. (1994). *Behavioral Expressions and Biosocial Bases of Sensation Seeking*. Cambridge University Press. Doi: 10.2277/0521432006

Zuckerman, M. (2007). Sensation Seeking and Risk. In M. Zuckerman, *Sensation seeking and risky behavior*. (pp. 51–72). Washington: American Psychological Association. Retrieved from http://content.apa.org/books/11555-002

Combining General And Specific Measures of Risk Preference Boosts Predictive Power in the Wild

Supplemental Material

Schürmann, O., Frey, R., Hertwig, R., Rieskamp, J., & Pedroni, A.

This document contains detailed supplemental material for the manuscript "Combining General and Specific Risk Preferences Measures Boosts Predictive Power in the Wild". These detailed explanations of some additional methods and results enhance the readers understanding of the main findings in the article. They are, however, not essential for understanding the article. The material might be of interest to experts in the field and for replication purposes.

Supplemental Material A: Selection of Preference Measures

We selected 4 measures that have previously been associated with risk taking in traffic from the original study of Frey et al. 2017.

Impulsivity (IMP)

The Barratt Impulsivity Scale (Barratt, 1965) measures the tendency to act without forethought (Lauriola, Panno, Levin, & Lejuez, 2013).
Forethought is considered an essential quality that marks a safe driver in traffic because it increases the perception of hazards on the road (Horswill, 2016). Impulsivity correlates with risky road-traffic behavior including dangerous street crossing (Schwebel, Gaines, & Severson, 2008). The measure is predictive for risk taking far beyond the traffic domain Lauriola et al. (2013). Consequently, we classify it to be a relatively domain-general measure.

Balloon Analogue Risk Task (BART)

The Balloon Analogue Risk Task is a computerized behavioral task (Lejuez et al., 2002). The BART and its predecessor have predicted risky street crossing of children (Hoffrage, Weber, Hertwig, & Chase, 2003), crash injury risk behavior (Vaca et al., 2013), or self-reported risky driving (Ba, Zhang, Peng, Salvendy, & Crundall, 2016). The BART was developed to capture risk preference in general without a specific focus on one domain and generalizes to domains such as substance abuse, gambling or unprotected sex (see (Lauriola et al., 2013). We classify it as a relatively domain-

general measure.

Vienna Risk Taking Test – Traffic (VRTT)

Specific risk preference was measured with the Vienna Risk Taking Test – Traffic (VRTT) (Hergovich, Arendasy, Sommer, & Bognar, 2007). The risk score has been found to correlate with naturalistic risk-taking behaviors in the street and is widely used in risk-assessment centers for road traffic (see Vingilis et al., 2015). We classify it as a relatively domain-specific risk preference measure.

Supplemental Material B: Confounding Variables

Due to the complexity of our real-life experiment, we tested the influence of confounding variables (i.e. weather conditions, time difference between the two sessions, gender, age, hours of walking in traffic per week) onto risk-taking in the street crossing task with Bayesian multiple linear regression models. None of the confounding variables or combinations thereof indicated a substantial influence on to risk-taking in the street crossing task (BF_0 : 0.06 − 0.9)

Supplemental Material C: Cognitive Modeling of Behavior in the BART

The Bayesian Sequential Risk-Taking Model (BSR)

BSR Model 3. The BSR3 was the best fitting model in the original study of Wallsten, Pleskac, and Lejuez (2005) and assumes that the decision maker has an initial belief about the probability $\hat{q}$ that balloon h will not explode on any given pump. This belief is modeled with a beta distribution over $\hat{q}$, with $a_h > 0$ and $b_h > 0$, estimated from the data (Pleskac, Wallsten, Wang, & Lejuez, 2008):

$$q = \frac{a_h}{a_h + b_h}, \tag{1}$$

The variance of the distribution, indicating the decision maker's uncertainty in the

initial belief, is determined as,

$$\delta = \frac{a_h b_h}{(a_h + b_h)^2 (a_h + b_h + 1)} \tag{2}$$

The expected gain of pumping balloon h at opportunity i equals

$$v_{h,i} = (q_h)^i (ix)^{\gamma+} \tag{3}$$

where q_h is the probability that balloon h will not explode after i pumps, and x is the reward for each successful pump. From the probability and the reward magnitudes, a target number of pumps is defined as the maximum of Equation 4 (Wallsten et al., 2005), which is

$$G_h = \frac{-\gamma^+}{\ln(q_h)} \ . \tag{4}$$

Finally, the BSR models how reward evaluation and learning experience are translated into the probability r_i that the balloon is pumped over time using a logarithmic choice function:

$$r_{h,i} = \frac{e^{\beta b_{h,i}}}{1 + e^{\beta b_{h,i}}} \tag{5}$$

where β is a free parameter representing how consistently participants follow their target number of pumps (i.e.,G_h). $d_h(i)$ is the distance at opportunity i from the targeted number of pumps, $d_h(\mathrm{i}) = i - G_h$. In BSR3, the participant evaluates how many pumps the target is before every trial.

BSR Model 1. The BSR1, the second best fitting model of the original study of Wallsten et al. (2005), assumes the participant evaluates the options of pumping or

stopping *before every pump opportunity*. Thus, the decision maker evaluates the utility of pumping for each balloon h on each pumping opportunity i as

$$b_{h,i} = (1 - p_{h,i})x - p_{h,i}\theta((i-1)x)^{\gamma-} \quad (6)$$

with $p_{h,i}$ denoting the probability of balloon's bursting given $(i - 1)$ successful pumps. The free parameters γ^- and θ are estimated from the data. This evaluation is then fed into a logistic response model to calculate the probability $r_{h,i}$ of the DM's pumping (see equation 5).

with β being a free parameter representing choice consistency.

BSR with fixed behavioral consistency parameters. For both models (BSR1 and BSR3) we created an additional alternative version (BSR1b and BSR3b), where we fixed the β parameter. The fixed value was derived from the mean β of the original BSR1 and BSR3 model versions. This was done due to previously reported inter-correlations of the model parameters.

Model Fitting Procedures

To estimate the free parameters of each of the BSR models we used the techniques as suggested in the Wallsten et al. (2005) paper. See the original for further elaborations and explanations of the model and fitting procedures.

Results from the BSR Model Comparisons

Table S1 shows the fitting results for the different versions of the BSR model. BIC stands for Bayesian information criterion. Results suggest that the BSR1b version with fixed behavioral consistency parameter best fits the data (mean *BIC: 157.37*) of most participants. We therefore used the γ parameter of this model as a measure of risk taking preference.

Table S1

BSR model fitting comparisons

Model	*df*	No. of DMs best fit	Mean BIC	Mean MLL
$BSR1_a$	5	0	158.63	-70.81
$BSR1_b$	4	20	157.37	-72.50
$BSR3_a$	4	15	157.94	-72.175
$BSR3_b$	3	15	168.58	-79.19

Note: Dm = decision maker; BIC = Bayesian information criterion; MLL =maximum log likelihood.

Supplemental Material D: Results of Risk Preference Combination Models

We implemented two different approaches to investigate the predictions of risk-preference combination models; A Bayes Factor analysis and a classical stepwise regression analysis. Figure S1 shows BF_0 for all combinations of linear models we tested, including all interaction terms possible. Table S2 shows the results of the classical hierarchical multiple linear regression analysis of the best fitting model. In step one, we entered the VRTT and SS as regressors with the SCT as dependent variable, and in step two we entered the BART. Results indicate that the BART (β=.168,t= -1.706,p= .09) is neither a significant predictor after VRTT (β=.01,t= 3.824,p< .001) and SS (β=.01,t=-3.25,p=.002) are already entered, nor is the change in R2 (F (1.79)=2.91,p=.09).

Figure S1. Bayes Factors of combination models predicting street crossing

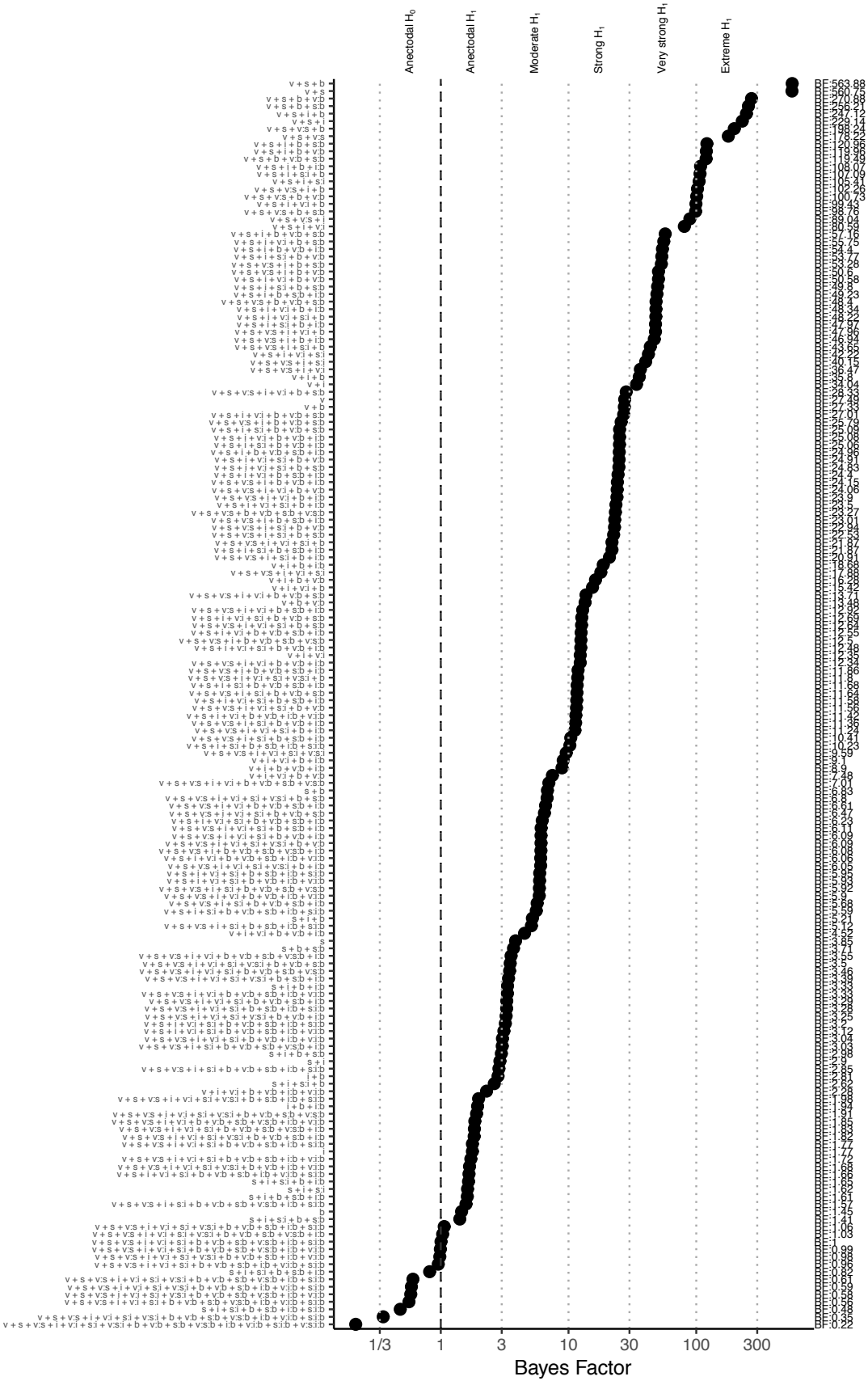

Table S2

Classical stepwise linear regression analysis for the best fitting models

Step 1

Term	β	95% CI	t	df	p
Intercept	-1.18	[1.91, 3.12]	-0.82	47	.41
VRTT	0.05	[0.03, 0.06]	4.05	47	$<.001$
SS	-0.05	[-0.07, -0.04]	-3.28	47	.002

Step 2

Intercept	-1.09	[-3.93, 1.75]	-0.77	46	.444
VRTT	0.04	[0.02, 0.07]	3.82	46	$<.001$
SS	-0.05	[-0.08, -0.02]	-3.26	46	.002
BART	-0.29	[-0.62, 0.05]	-1.71	46	.095

Note: Explained Variance for step 1: $R^2 = .34$, for step 2: $R^2 = .39$

References

Ba, Y., Zhang, W., Peng, Q., Salvendy, G., & Crundall, D. (2016, January). Risk-taking on the road and in the mind: behavioural and neural patterns of decision making between risky and safe drivers. *Ergonomics*, *59*(1), 27–38. doi: 10.1080/00140139.2015.1056236

Hergovich, A., Arendasy, M. E., Sommer, M., & Bognar, B. (2007, July). The Vienna Risk-Taking Test - Traffic: A new measure of road traffic risk-taking. *Journal of Individual Differences*, *28*(4), 198–204. doi: 10.1027/1614-0001.28.4.198

Hoffrage, U., Weber, A., Hertwig, R., & Chase, V. M. (2003). How to keep children safe in traffic: Find the daredevils early. *Journal of Experimental Psychology: Applied*, *9*(4), 249–260. doi: 10.1037/1076-898X.9.4.249

Horswill, M. S. (2016). Hazard perception in driving. *Current Directions in Psychological Science*, *25*(6), 425–430.

Lauriola, M., Panno, A., Levin, I. P., & Lejuez, C. W. (2013, January). Individual differences in risky decision making: A meta-analysis of sensation seeking and impulsivity with the balloon analogue risk task: personality and risky decision Making. *Journal of Behavioral Decision Making*, *27*(1), 20–36. doi: 10.1002/bdm.1784

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84. doi: 10.1037//1076-898X.8.2.75

Schwebel, D. C., Gaines, J., & Severson, J. (2008, July). Validation of virtual reality as a tool to understand and prevent child pedestrian injury. *Accident Analysis & Prevention*, *40*(4), 1394–1400. doi: 10.1016/j.aap.2008.03.005

Vaca, F. E., Walthall, J. M., Ryan, S., Moriarty-Daley, A., Riera, A., Crowley, M. J., &

Mayes, L. C. (2013). Adolescent Balloon Analog Risk Task and behaviors that influence risk of motor vehicle crash injury. *Annals of advances in automotive medicine*, *57*, 77.

Vingilis, E., Roseborough, J. E., Wiesenthal, D. L., Vingilis-Jaremko, L., Nuzzo, V., Fischer, P., & Mann, R. E. (2015, February). Experimental examination of the effects of televised motor vehicle commercials on risk-positive attitudes, emotions and risky driving inclinations. *Accident Analysis & Prevention*, *75*, 86–92. doi: 10.1016/j.aap.2014.11.008

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological Review*, *112*(4), 862–880. doi: 10.1037/0033-295X.112.4.862

Risk perceptions: Their impact on risk taking and the effect of early experience

Oliver Schürmann

University of Basel

Timothy J. Pleskac

Max Planck Institute for Human Development, Berlin

Renato Frey

University of Basel and Max Planck Institute for Human Development, Berlin

Author Note

This research was supported by the Swiss National Science Foundation with a grant to the first author (P0BSP1_148884). We would like to thank Jann Wäscher and Chantal Wysocki for assistance with data collection. We also thank Laura Wiles for editing the manuscript. Finally, we thank Andreas Pedroni and Ralph Hertwig for comments in the development of this project. Author contributions were as follows: Conceptualization: O.S., R.F., & T.J.P.; Methodology: O.S., & T.J.P.; Software: O.S.; Data collection & curation: O.S.; Formal analysis: O.S.; Writing - Original Draft: O.S.; Writing - Reviewing & Editing: O.S., R.F., & T.J.P; Supervision:T.J.P. The full data set as well as the scripts for the statistical analysis can be downloaded here (Annonymous Link): Correspondence concerning this article should be addressed to Oliver Schürmann, University of Basel, Missionsstrasse 62A, 4055 Basel, Switzerland. E-mail: o.schuermann@unibas.ch.

Abstract

What are the properties of people's perceptions of the risks involved in the decision they face? And what role do these perceptions play in the risks people ultimately take? We investigate these questions using the Balloon Analogue Risk Task (BART), a widely used laboratory risk-taking task. In the BART, participants inflate a virtual balloon, earning points if the balloon does not explode. An effective solution to performing this task is for participants have to gauge the probability of the balloon exploding and use this perceptions of the risks to determine to pump or not. To date, how people perceive risks in this task as well as the factors driving risk perception remain unclear. Thus, across two experiments we examined risk perceptions with self-reported probability-ratings. Our results show that the probability ratings deviate not only from the actual probabilities, but also from those predicted by the most successful cognitive models of the BART. Yet, risk perception ratings correlated with actual choice behavior. Moreover, risk perceptions in the task mediate the influence of early initial experiences on the amount of risk participants are willing to engage in. Our results highlight the important role of risk perception in decision processes in which uncertainty about the choice options prevails.

Risk perceptions: Their impact on risk taking and the effect of early experience

A critical determinant of whether people engage in risk-taking activities is how they perceive the chance of a loss (Slovic, 1964). Individual perceptions of probabilities can help explain differences in risk taking between people, cultures, and domains (Gigerenzer, 2006; Klos, Weber, & Weber, 2005; Siegrist, Keller, & Kiers, 2005; Viscusi, 1990; Weber, Blais, & Betz, 2002; Weber & Hsee, 1998).[1] This linkage between risk perceptions and risk-taking behavior has typically been investigated by means of self-reports about past behaviors (e.g., smoking)(Viscusi, 1990) or scenarios involving hypothetical behaviors (Weber & Hsee, 1998). The practice of using self-reported risk-taking behavior reflects a common method of studying risk-taking behavior in general (Frey, Pedroni, Mata, Rieskamp, & Hertwig, in press; Jackson, Hourany, & Vidmar, 1972; Weber et al., 2002; Zuckerman, 2007). The reason for their popularity is that despite the criticism that self-reports may prompt socially desirable responses (Crowne & Marlowe, 1960; Paulhus, 1984), they also have several advantages. On the one hand, they easily permit assessing a wide range of behaviors. On the other hand, several studies have found that specific self-report measures exhibit desirable test-theoretic criteria (Weber et al., 2002; Zuckerman, 2007), and recently a comprehensive assessment of a battery of 39 measures found self-report measures to have substantial convergent validity and test–retest reliability (Frey et al., in press).

Yet, whereas self-reports have also been used to examine the cognitive processes underlying risky decision making (Weber et al., 2002), they might be somewhat limited in the depth of mechanistic explanations they can provide—a problem that behavioral tasks promised to overcome. Cognitive psychologists interested in risk-taking behavior have thus often turned to laboratory-based gambling tasks (e.g., Bechara, Damasio, Damasio, & Anderson, 1994; Figner, Mackinlay, Wilkening, & Weber, 2009; Frey, Rieskamp, & Hertwig, 2015; Lejuez et al., 2002; Pleskac, 2008; Pleskac, Wallsten, Wang, & Lejuez, 2008; Rogers et al., 1999; Slovic, 1966). In these tasks, participants

[1] Risk perception may refer to different aspects of risky decisions; in this paper we focus on the perception of probabilities.

repeatedly have the option of taking a risk (typically for real payoffs) in a controlled laboratory environment. A popular task, for instance, is the Balloon Analogue Risk Task (BART; Lauriola, Panno, Levin, & Lejuez, 2013; Lejuez et al., 2002; Pleskac et al., 2008). During the BART, participants are given the option of pumping a virtual balloon shown on a computer screen. Each successful pump pays participants a small payoff (e.g., 5 cents). However, participants are told that somewhere between the first pump and when the balloon fills the screen, the balloon will explode. If the balloon explodes, they lose all their earnings for that trial. Participants can stop pumping the balloon and collect their earnings to prevent this loss. If they choose to stop and collect their earnings, the trial ends and the earnings are transferred to a permanent bank.

Participants in the BART typically complete a series of trials (e.g., 30) and the average number of pumps on non-exploding balloons across these trials is used as a general risk-taking score (the adjusted BART score). The adjusted BART score has been found to correlate with real-world risky behaviors including substance use and abuse, delinquency and risky sexual behaviors (e.g., Aklin, Lejuez, Zvolensky, Kahler, & Gwadz, 2005; Hopko et al., 2006; Lejuez et al., 2007; Skeel, Pilarski, Pytlak, & Neudecker, 2008). This relationship to risk-taking behaviors has opened up the possibility of using the BART and similar tasks to study the processes underlying risk-taking behavior in a controlled setting both at the cognitive (Pleskac, 2008; Rolison, Hanoch, & Wood, 2012; Wallsten, Pleskac, & Lejuez, 2005) and neural level (Helfinstein et al., 2014; Rao, Korczykowski, Pluta, Hoang, & Detre, 2008; Schonberg et al., 2012). In this article, we build on this approach and study risk perceptions in the BART with the aim of better understanding some basic properties of risk perceptions and their link to risky behavior.

One reason why the BART is particularly interesting for this goal is that the stochastic properties of the task are well defined. At the same time participants are not instructed about the underlying properties of the task. This means we can use the BART to examine the correspondence between people's risk perceptions and the true structure of the task. In fact, with the goal of mimicking many risks outside the lab

that increase sequentially (e.g., smoking), the probability of an explosion in the BART increases with each pump (Lejuez et al., 2002). The explosion points (across trials) are predefined by a random draw of numbers without replacement between 1 and a maximum number of pumps possible n (usually 128). Thus, the *a priori* probability of a balloon exploding on pump i, given that it did not explode on the preceding pumps, is

$$p = \frac{1}{n - i + 1}, \tag{1}$$

where n is the number of maximum pumps that can be made (in most cases 128) and i is the size of the balloon at the decision stage in a trial (i.e., pump opportunity). So, with a maximum of 128 pump opportunities, the probability that the balloon would explode on the first pump would be $\frac{1}{128}$. Given that it did not explode on the first pump, the probability on the second pump opportunity would be $\frac{1}{127}$, on the third pump opportunity, $\frac{1}{126}$ and so forth up to pump opportunity 128 where the balloon would explode with certainty. The solid line in Figure 1 plots the probability of a balloon exploding given its current size (i.e., the conditional probability) for every possible pump opportunity i.

Because participants are not explicitly informed about the probabilities in the BART, an important question is to what extent their risk perceptions track these objective probabilities. One answer to this question comes from the Bayesian Sequential Risk Taking (BSR) model, a cognitive model developed to explain the decision process participants use in deciding how far to pump the balloon (Pleskac, 2008; Pleskac & Wershbale, 2014; Wallsten et al., 2005). According to the BSR, at the beginning of each trial, participants evaluate the possible pump options by trading off the probability of an explosion of the balloon (and losing the money in the bank) with the gain of successfully pumping the balloon for each possible trial. As a result of this evaluation process, participants identify the target pump that has the greatest expected utility for them. They then probabilistically pump to this target pump. Each trial thus ends with either a stop decision, centered on average around the target, or an explosion. Either way, participants update their beliefs about the explosion probability of the balloon and use this updated belief to estimate the expected utility of each pump on the next trial.

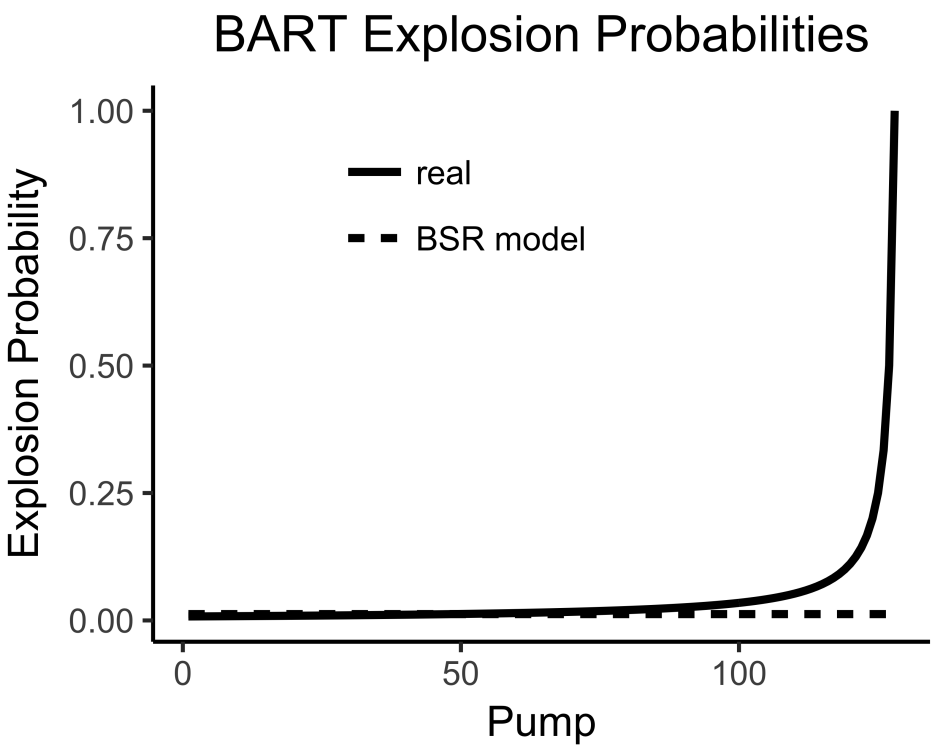

Figure 1. The solid line represents the underlying probability of an explosion for each of the 30 possible pumps in the BART, given that the balloon was already at the size (conditional probability). The dotted line represents the probability of an explosion per pump that the best-fitting cognitive model assumes on the first pump (median over all participants).

According to the best-fitting version of the BSR, which assumes that the probability of an explosion on any given trial is constant over all possible pumps, participants thus misunderstand the task and treat the non-stationary probability of an explosion as a stationary event.[2] The dotted line in figure 1 plots the median subjective probability of participants at the beginning of the task estimated from the data in the study of Wallsten et al. (2005). This estimate is inferred from choice behavior in the task. What has not been studied, however, is how well participants' reported perceptions track the

[2]The BSR with a stationary representation of probabilities, at least in the BART, accounts for choice behavior of most participants better than a variation of the model where participants have an accurate representation of the probabilistic structure of the task, but are uncertain about the actual number of maximal pumps, n, and learn this number over time.

explosion probabilities.

This leaves a number of unanswered questions: Do people's risk perceptions about the probabilities match the structure of the BART and show an increase in the probability of an explosion with increasing pumps? Or are they constant, as the BSR model predicts? Are their risk perceptions associated with their pumping behavior? And finally, what are determinants of their risk perceptions? In two studies, we investigated how participants rate the probability of an explosion in the BART and how these judgments influence their behavior in the task.

In the first study, we asked participants to complete the standard BART and then to rate the probability of an explosion for different sizes of the balloon. A complementary goal of this study was to replicate the results found by Lejuez, Aklin, Jones, et al. (2003), namely, that smokers were more risk seeking in the BART than non-smokers. To foreshadow our results, although we did not find a difference in pumping behavior between smokers and non-smokers, we did find some intriguing results in terms of participants' risk perceptions. A second study followed up on these results and investigated how initial experiences shape participants' risk perceptions, and how these perceptions change over the course of the task.

Study 1: Properties of Risk Perceptions in the BART

To investigate risk perception in the BART, we repeated one of the original BART studies, which found smokers and non-smokers to differ in terms of the adjusted BART score (Lejuez, Aklin, Jones, et al., 2003). We thus invited a group of smokers and a group of non-smokers to undertake the classic version of the task. After the task, we asked participants to sequentially rate the probabilities of an explosion after a certain amount of pumps for several balloon sizes. We hypothesized that people differ substantially in their perception of the probability and that this difference is connected with behavior in the task.

Methods

Participants. We invited 100 participants ($M_{age} = 25$ years, $SD_{age} = 4$ years, 59% female) from the participant pool of the Center for Adaptive Rationality at the Max Planck Institute for Human Development. We required half of them ($n = 50$) to have been daily smokers (smoking at least one cigarette per day over the past six months) and the other half to be non-smokers (no history of daily smoking). The sample size was determined using a power analysis based on the mean correlation between the adjusted BART score and smoking status (*mean* $r = .33$), and a desired power of .9. These parameters result in a goal of 86 participants. We rounded this number to 100. The groups were balanced in gender (smokers: $F = 56\%$, $M = 44\%$; non-smokers: $F = 62\%$, $M = 38\%$). The participants had no prior knowledge of the experimental procedures and had no previous encounter with the BART. All of the participants were native German speakers. Each participant signed a consent form prior to the beginning of the experiment. Participants were paid €10 per hour for their participation in the study plus a bonus contingent on their performance in the BART. The study was approved by the Institutional Review Board of the Max Planck Institute for Human Development.

Materials. All tasks were conducted on personal computers with 19-inch screens. The first part of the study (behavioral tasks) was presented using the *Presentation* software from Neurobehavioral Systems (Albany, California). The second part of the study was presented in Unipark, an online questionnaire tool.

Balloon Analogue Risk Task. During this computer-controlled task, participants were presented with a series of balloons on the computer screen (i.e., trials). For each trial, participants could sequentially click a button to pump up the balloon. For each successful pump, €0.05 was added to a temporary account. At any point, the participant could choose to stop pumping (i.e., to "cash out") whereupon a slot machine payoff sound was played and the money in the temporary account was transferred to a safe account. However, if the balloon exceeded a certain number of pumps, it exploded, with the computer playing a pop sound, and the money in the

temporary account for this trial was lost. In the original BART, the balloon could fill up the entire screen (Lejuez, Aklin, Jones, et al. (2003)). Due to different monitor types, we placed the balloon within a square on the screen and the balloon could only get as large as the square.

A new balloon appeared after each trial until a total of 30 balloons was completed. During the trial, the screen showed five items: the balloon, the amount won in the last trial, the total amount won (i.e., the safe account) and two buttons "pump" and "cash". The sound files and amount of information on the screen were the same as used in the original study of Lejuez, Aklin, Jones, et al. (2003). The instructions were translated to German from the original task.

In each of the 30 trials, the balloon was programmed to explode at a certain explosion point, which was fixed beforehand. The explosion points are listed in the supplemental material (see Table 1). For half the participants, we used the same series of explosion points as used in the original study of Lejuez, Aklin, Jones, et al. (2003). For the other half of the participants, we inverted this order by subtracting every explosion point of the original series from 128. Consistent with conventional BART procedures (Lejuez, Aklin, Zvolensky, & Pedulla, 2003), participants received no detailed information about the probability of an explosion. They were told that the balloon could explode anywhere between the first pump and when the balloon reached the size of the box surrounding the balloon. At the end of the task, a trial was randomly selected and participants were paid according to their earnings on that trial

Probability Ratings. To measure risk perceptions, we asked participants to rate the probabilities of explosions in the task. Participants saw a sequential series of 11 balloons at different pump stages (i.e.,a sizes). The series of sizes ranged from 1 pump to 128 pumps in 11 equal-sized steps: $0, 13, 26, 39, 52, 65, 78, 91, 104, 117,$ and 128. At each step participants were prompted with the question (in German), "What do you think is the probability that the balloon will explode with one additional pump, given it is already at this size?" Participants rated the probability from 1 to 100% using a slider below the balloon.

Questionnaires and Additional Measures. In addition to *gender* and *age*, we also assessed *smoking status* using the Fagerström Test for Nicotine Dependence and *alcohol dependence* using the Alcohol Use Disorders Identification Test (Saunders, Aasland, Babor, de la Fuente, & Grant, 1993). We also included two personality measures to capture *sensation seeking* (Zuckerman, 2007) and *impulsivity* (Barratt, 1965). Those two traits have repeatedly been shown to be correlated with real-life risk-taking behavior in various domains such as drug use, financial decisoins, and health decisions including smoking (see Lauriola et al., 2013, for a review). Both measures were also included in the original study of Lejuez, Aklin, Jones, et al. (2003).

In addition to the tasks above, we also had participants to complete two other tasks that we thought might complement the risk perception task. One task, the *benefits task*, was designed to measure individual differences in participants' benefits from pumping a balloon at different pump stages. The other measure, the *BART lottery task*, was a descriptive lottery-based gambling task that mimicked some of the stochastic properties of the BART. Both measures resulted in unreliable estimates that seemed not to be linked to behavior in the BART, which is why we do not report them further. Please consult the supplemental material for details about the methods, results, and interpretations of the BART benefit and lottery tasks.

Procedures. After signing an informed consent form, participants completed the two parts of the study. In the first part, participants completed the behavioral tasks on the computer, including the BART, probability rating task, benefit task, and lottery task, in that order. As the experiment took part in a laboratory where participants were not in an enclosed cubicle, we asked them to wear headphones during the first part of the study to help reduce influence by the other participants' clicking during the tasks.

In the second part of the experiment, participants completed three questionnaires including the personality scales sensation-seeking and impulsivity and a questionnaire about the participant's demographic information. The questionnaires were filled out in the online questionnaire program unipark. After finishing all tasks, participants were paid €12 per hour and an additional bonus depending on their performance in the

BART and the lottery task, ranging from €3 to €6.

We should point out that there are some deviations between our study and the study of Lejuez, Aklin, Jones, et al. (2003). In our experiment the BART was administered only once while in the original study, the BART was administered three times. Because Lejuez, Aklin, Jones, et al. (2003) did not find any significant differences between the three administrations, there was only one run of the BART in our study. The original study did not include the BART perception, benefit and lottery task, but did include the Iowa Gambling Task (IGT; Bechara et al., 1994). Because the IGT was not the focus in our study, we did not include this task. Furthermore, (Lejuez, Aklin, Jones, et al., 2003) only used 34 non-smokers and 26 smokers. The sample in the original study included English-speaking undergraduate students who were younger than in our sample (mean age: 20.1 years), whereas our participants were German speaking and drawn from a participant pool made up primarily of undergraduate students from a local university (average age of 25.1 years). Finally, in our study, participants were incentivised by receiving the winnings of random balloon draw in the BART. In the original study, participants were told they would win money depending on how well they performed compared to other participants (a total score in the bottom third earned $5, a total score in the middle third earned $10, and a total score in the top third earned $15.

Analysis. With regard to the probability ratings, we adapted a psychometric function (henceforth rating function) to characterize the individual ratings of each participant. The function function is given by:

$$p\left(x\right)_i \ = \gamma_i \ + (1 - \gamma_i) \ \cdot \frac{1}{1 + e^{-\frac{x - \mu_i}{\theta_i}}}, \tag{2}$$

with $p(x)_i$ denoting the subjective probability of the balloon exploding on pump x for participant i. The parameters γ, μ, θ are free parameters estimated individually from the data for each participant.

Figure 2 illustrates how the different parameters impact the function. The μ parameter (threshold parameter) estimates at what pump the probability of an

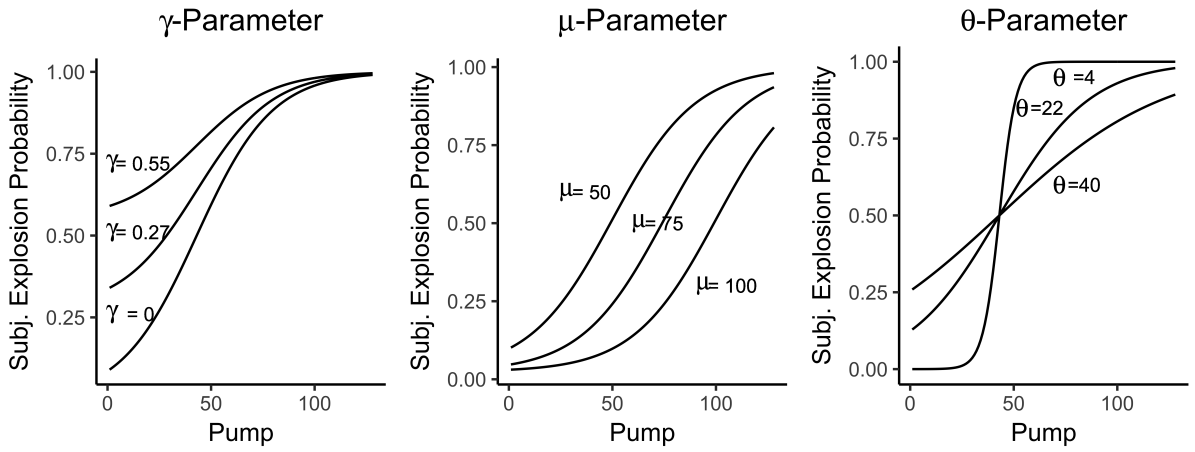

Figure 2. The figure shows how the rating function, (Equation 2) which is used to characterize the probability ratings, is impacted by the different parameters. Note as shown in the leftmost panel, the function can closely align with the actual probabilities in the BART (see Figure 1.)

explosion was .5. Thus, the higher μ was, the lower was the probability rating over all 11 balloon sizes probed. The θ parameter (sensitivity parameter) controls the slope of the function. Pragmatically, it indicates how quickly participants switched from low probability ratings to high probability ratings. Finally, the γ parameter (height parameter) measures the starting point of participants' probability ratings. We fitted a separate curve to each participant's probability ratings by minimizing the squared deviations between the observed ratings and those predicted by the curve. For this, we used the optimize algorithm in R, with the following boundaries for the parameter estimation: $\mu : [-128, 128]$, $\theta : [-1, 200]$, $\gamma : [0, 1]$.

Results

BART Results. On average, participants had an adjusted BART score (mean number of pumps in all trials where the balloon did not explode) of 37.5 ($SD = 12.8$). This is consistent with the average score of 37.6 ($SD = 11.6$) reported in Lejuez, Aklin, Jones, et al. (2003). However, unlike in the original study, smokers ($M = 36.83$; $SD = 13.86$) did not significantly differ from non-smokers ($M = 38.2$; $SD = 11.93$) in the adjusted BART score ($F(1, 98) = 0.28$, $d = 0.10$, $p = .60$). We thus did not replicate

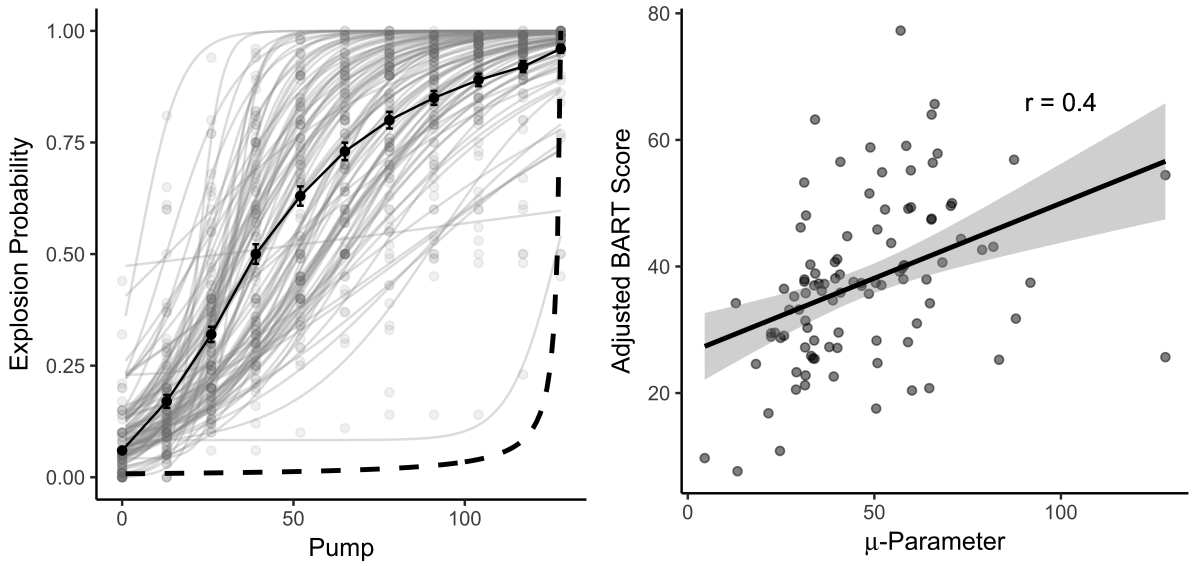

Figure 3. Subjective probability ratings in Study 1. **Left Panel:** Individual probability ratings for each balloon size in the perception task and the fitted rating function curves (light gray). The individual probability ratings generally increased with balloon size and the mean of the ratings for each size (solid curve). The estimates are strikingly different from the objective probabilities of the task (dashed curve). Error bars on the mean ratings are ±1 standard error. **Right Panel:** The relationship between the individual μ parameter values (thereshhold parameter) and the adjusted BART score.

the findings of Lejuez, Aklin, Jones, et al. (2003) concerning the differentiation of smokers and non-smokers with the BART. Means and standard deviations for all measures as well as correlations among all measures risk-variables are provided in the supplemental material.

Probability Perception Task Ratings. The left panel of Figure 3 shows the probability ratings of each participant (light gray dots), the fitted function over each of these dots (light gray curves), the mean rating for each of the 11 rated balloon sizes (black dots), and the fitted function to the average rating (black curve). On average, participants' estimates of the probability of an explosion did increase with each pump. The ratings for each pump, however, were all substantially higher than the true probability (dashed curve). The rating function fitted the subjective probability ratings of participants well (mean R^2 = .96).

Table 1

Correlations between rating function parameters and adjusted BART score in Study 1

	M	SD	μ	θ	γ	adj. BART
Threshold μ	47.36	21.39		0.55***	0.35***	0.40***
Sensitivity θ	19.76	20.12			0.31**	0.00
Height γ	0.02	0.04				-0.12
adj. BART	37.52	12.79				

Note. p < .001 ***, p < .01 **, p < .05 *

Table 1 shows correlations among the individual parameter values of the rating function as well as correlations with the adjusted BART score. The threshold parameter (horizontal shift of the curve) μ correlates with the adjusted BART score $r = .40$, 95%CI [.22, .55], $t(97) = 5.23$, $p < .001$), whereas the θ parameter ($r = .00$, 95% CI [−.20, .20], $t(98) = 0.01$, $p = .989$) and the γ parameter did not significantly correlate with the adjusted BART score ($r = -.12$, 95% CI [−.31, .08], $t(98) = -1.16$, $p = .248$). The right panel of Figure 3 shows the scatter plot of individual μ parameter values and the average number of pumps in the task. Thus, higher pumping was correlated with the threshold (i.e. the pump) on which participants rated the explosion probability to be 50%. In other words, participants with lower explosion probability ratings also pumped more in the task. This result indicates that individual differences in risk taking in the BART task (i.e., the number of pumps taken on non-exploded balloons) can be attributed to risk perceptions in the task. Note that smokers and non-smokers did not differ in any of the probability-rating parameters in the BART (threshold μ: $F(1, 98) = 0.35$, $p = .555$, $\eta^2_G = .004$; sensitivity γ: $F(1, 98) = 1.36$, $p = .247$, $\eta^2_G = .014$; height γ: $F(1, 98) = 0.20$, $p = .655$, $\eta^2_G = .002$.

Discussion

With the first study we aimed two answer two questions; How can perceptions of probabilities be represented, and how are they associated with behavior? Addressing the first question, Study 1 showed that when people are asked about their perceptions of probabilities in the BART, the results differ from previous assumptions. The ratings were generally a lot higher than the actual probabilities and showed a different functional form than the actual hyperbolic function. Perhaps more importantly, the form was quite different than that assumed by the cognitive models of the BART (Pleskac, 2008; Wallsten et al., 2005), which would seem to call into question their psychological plausibility. The change in the ratings over increasing balloon size could be well described by an s-shaped rating function.

Regarding the second question and the relationship between perceptions of probabilities and behavior, the ratings were related to risk-taking behavior, with participants who gave lower probability estimates (low μ) pumping more. However, this relationship is perhaps not all that surprising given that the ratings were collected after participants had completed the BART. To better investigate the nature of the connection between risk perception and risk taking in our experiment, we conducted a second study to see whether we could manipulate the perception of explosion probabilities in the early stages of the task and whether this caused a change in behavior.

Study 2 - The Impact of Early Experience on Risk Perceptions and Risk-Taking Behavior

In Study 2, we sought to assess how well participants' risk perceptions predicted their risk-taking behavior. To this end, we collected risk perceptions at two time points: after the first balloon and after the last balloon. Doing so also provided us a means to test a hypothesis that initial experiences in the BART play a large determining role in participants' risk perceptions, which in turn impact risk-taking behavior throughout the task.

Our hypothesis is consistent with several recent findings. For one, in similar experience-based tasks, the first experienced payoff has a substantial and lasting effect on participants' subsequent choice behavior (Shteingart, Neiman, & Loewenstein, 2013). Specifically, the first experienced payoff appears to disproportionately determine how participants value the alternative, a phenomena Shteingart et al. (2013) called outcome primacy. Second, in the BART, Koscielniak, Rydzewska, and Sedek (2016) showed that the explosion points over the first three balloons impact pumping behavior throughout the task, with participants who experienced explosion points after the first few pumps pumping substantially less throughout the task than those for whom the explosion points were near the end (i.e., when the balloon filled the screen).[3] Similarly, Walasek, Wright, and Rakow (2014) showed that variability in the explosion points (while keeping the average explosion point constant) over the first 10 balloon trials led to participants having greater confidence in their estimates of the value of a certain balloon.

Our hypothesis would explain the effect of early experiences on risk-taking behavior in the BART based on people's risk perceptions. That is, we propose that risk perceptions play a mediating role between participants' early experiences and risk-taking behavior in the BART. To test this hypothesis, we took the Koscielniak et al. (2016) study one step further and manipulated only the explosion point of the first balloon, holding everything else constant. Specifically, for the control group, we set the explosion of the first balloon at the same point as in the standard version (i.e., on pump 65). For the other two groups we set the explosion either earlier (i.e., pump 13) or later (i.e., pump 96), respectively.

[3]In a recent study, Wichary, Pachur, Koscielniak, Rydzewska, and Sedek (2017) used the BSR to show that in the Koscielniak et al. (2016) data those participants exposed to early explosion points appeared to set their initial beliefs lower and have lower uncertainty in those beliefs than participants who were exposed to later explosion points. However, the modeling seems misspecified as technically the beliefs in the BSR correspond to participants' beliefs before beginning the task. Nevertheless, despite this misspecification, these results would seem to support our conjecture that early experiences impact participants' risk perceptions.

Methods

Participants. We invited 90 participants ($M_{age} = 24.5 years$, $SD_{age} = 4.5 years$, 61% female) from the Center for Adaptive Rationality's participant pool. The participants were randomly assigned to one of three conditions. We did not specifically invite smokers or non-smokers. The participants had no prior knowledge of the experimental procedures and had no previous encounter with the BART. All of the participants were native German speakers. Each participant signed a consent form prior to the beginning of the experiment. Participants were paid €10 per hour for their participation in the study plus a bonus contingent on their performance.

Materials. We used a similar set of tasks as in Study 1. We note the differences here.

Balloon Analogue Risk Task. We used the same BART from Study 1 and used the same explosion points as the original version in Lejuez, Aklin, Jones, et al. (see the supplementary for the entire list of explosion points). However, we manipulated the first explosion point in the following way. In the *control* condition the first explosion point remained the same, set at 65. In the *low* condition the explosion point was set to 13 and in the *high* condition the explosion point was set to 98.

Probability Ratings. We used the same probability rating task as in Study 1, but participants completed the task twice once after the first balloon and once after completing all 30 balloons.

Questionnaires and Additional Measures. As in Study 1 we also asked participants to rate the expected benefits that would come from pumping the balloon (benefits task). The benefits task was the same as in Study 1. Even though responses in this task were not very reliable in Study 1, we still collected it to investigate if the unreliability remained. Nevertheless, the reliability remained. Please see the supplementary materials for further details.

As in Study 1, we also asked participants to complete a lottery task where participants were asked to make a series of choices between a monetary lottery and a fixed payoff that was meant to mimic the decisions participants face at each pump. The

goal with this task was to examine a possible description-experience gap with a task like the BART (Hertwig & Erev, 2009). This question, however, is beyond the scope of the paper. Thus, we describe the task and some of the results in the Supplementary Material.

Finally, following similar procedures with past studies with the BART, we also asked participants to also complete the sensation seeking scale, the impulsivity scale, and a demographic questionnaire. These measures were collected more to standardize procedures and as a result we do not report further on them.

Procedures. We used a similar procedure to Study 1. The only major deviation was that in Study 2, participants completed the risk task twice, once after the first balloon (pre-task) and then after completing all 30 balloons (post-task).

Analysis. We used a similar set of analyses for Study 2. By collecting risk perceptions at the beginning of the BART, we were able to examine to what degree the relationship between early experience in the BART and overall risk-taking behavior is mediated by their risk perceptions. To do so, we conducted a mediation analysis using the adjusted BART score as dependent variable, the explosion point of the respective condition as predictor variable, and probability ratings (using the individually fitted μ parameters for each person, Because this was the best predictor of behavior in the BART in Study 1) as a mediator. We used the mediation analysis procedure specified by Preacher and Kelley (2011). This allowed us to calculate the indirect effect size of the influence of perception as a mediator between the explosion point and the behavioral change in pumping. We used the MBESS package in R, which is suggested by Preacher and Kelley (2011) and is especially useful to calculate bias-corrected and accelerated (BCa) bootstrap confidence limits.

Results

BART results. Figure 4 shows the distributions of the adjusted BART scores for each of the three explosion groups. The control groups average pumping behavior ($M = 42.47$, $SD = 11.42$) was slightly higher than in Study 1. This was considerably

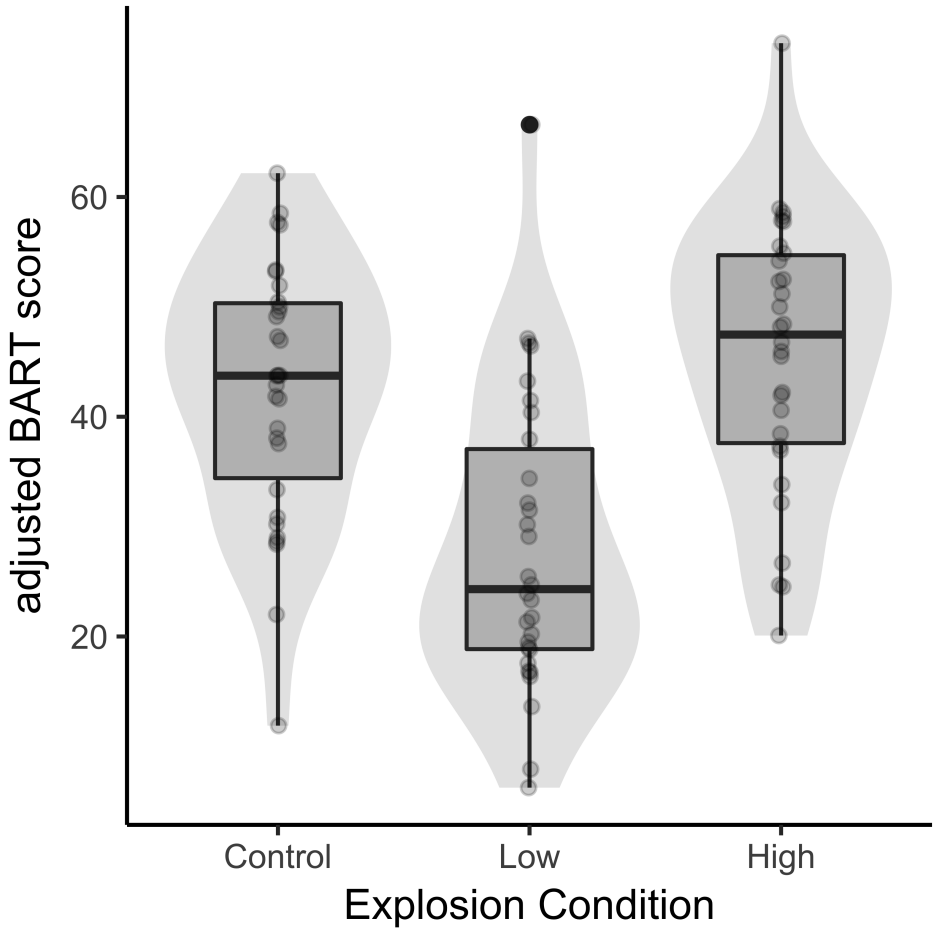

Figure 4. Adjusted BART scores for each condition (balloon size at explosion in trial 1) of the of the BART in Study 2. The differences between the conditions low and control as well as between low and high proved to be significant.

higher than in the low condition, where the balloon exploded very early ($M = 28.01$, $SD = 13.25$). The high condition group showed the highest average pumping behavior over all participants ($M = 45.67$, $SD = 12.47$). An ANOVA revealed a significant main effect of explosion condition on average adjusted BART score ($F(2, 87) = 16.71$, $p < .001$, $\eta^2_G = .278$) with the contrast between the low condition and the control group being significant ($t = 4.442, p < .001$). The difference between the control condition and the high condition was not significant ($t = 0.98, p < .33$). One possible explanation for the lack of a difference between the high explosion point group and the controls is that only seven participants (23%) in the high condition pumped the balloon until it exploded. In comparison, 11 participants (37%) in the control group and 25

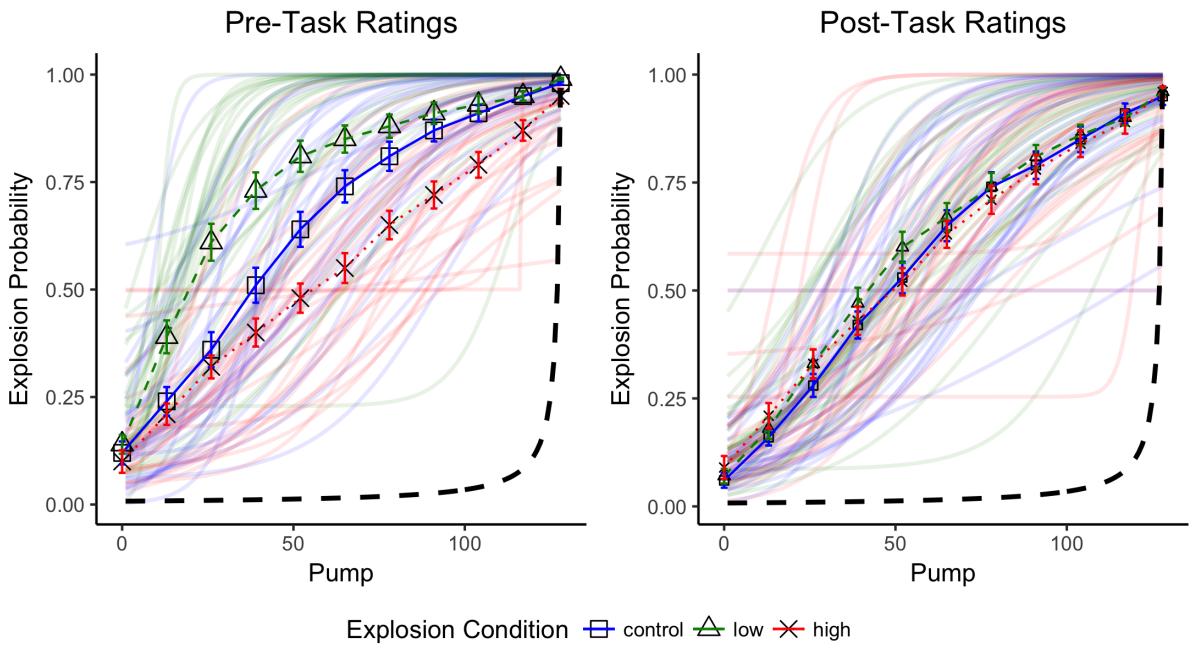

Figure 5. Rating functions fitted for each participant's probability ratings - green dashed curve (triangles): low condition; solid blue curve (squares) : control condition; dotted red curve (exes): high condition). The thick dashed curve shows actual probabilities in the task. **Left Panel:** Function curves (light colored) and the mean values for each of the 11 rating sizes separated by explosion condition in the pre-perception task. Error bars represent standard errors. Note the difference of the mean probability ratings for each condition. **Right Panel:** Fitted rating function curves (light gray) and the mean values for each of the 11 rating sizes separated by explosion condition in the post-perception task. Error bars represent standard errors.

participants (83%) in the low group experienced an explosion in the first trial. Further details about the descriptive results can be found in the supplemental material.

Probability Rating. Figure 5 shows the curves for the individually fitted rating functions for both the pre- (left) and the post-task ratings (right). The rating functions fitted the data well for both ratings, pre- and post-task (pre-task: mean $R^2 = 0.97$, post-task:mean $R^2 = 0.97$). The colors of the curves represent the three conditions (i.e., control: blue, low group: green, high group: red). The pre-task ratings differed substantially depending on the explosion condition. The low group (where the balloon exploded early) rated the probability that a balloon would explode generally

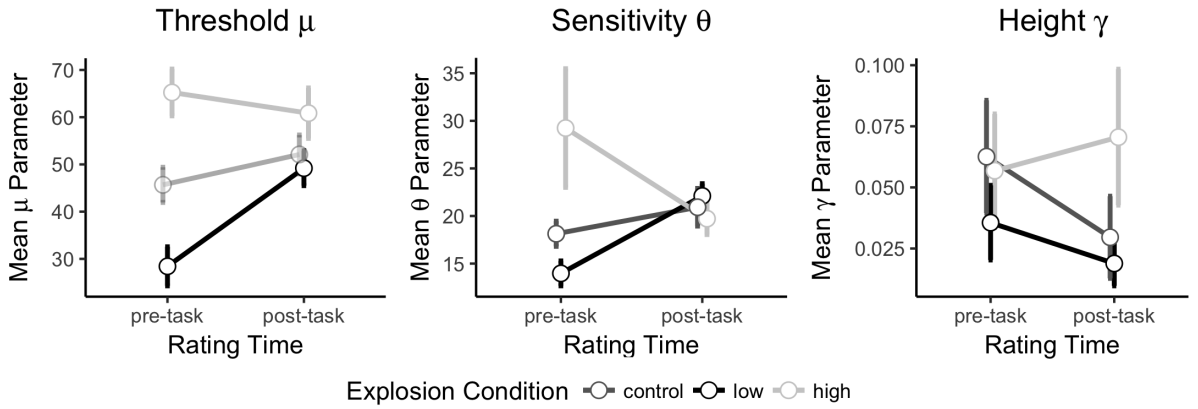

Figure 6. Mean parameter values of the individually fitted rating function by condition and rating time (pre-task: immediately after the first trial, post-task: after the last trial). **Left Panel:** mean γ-parameter values (shift of the rating function on the y-axis) per group and presentation time. **Middle Panel:** mean μ-parameter (shift of the rating function on the horizontal axis) values per group and presentation time. **Right Panel**: mean θ-parameter values (slope of the rating function) per group and presentation time

.

higher than in the control group (blue) or the high group (red). This difference is not prevalent in the post-task ratings. As in Study 1, the probability ratings generally increased over the size of the balloon and were substantially higher than the actual probability in the BART.

To investigate the influence of the explosion point on the probability ratings, we examined the parameters of the rating function separately. Figure 6 shows, for each parameter of the rating function, the mean per explosion condition and administration time (i.e., pre- and post-rating). We conducted a set of repeated-measures ANOVA with rating time as a repeated-measure variable. In each analysis we looked at the influence of the two factors time and condition as predictors, with contrasts for the condition between high explosion point and control and low explosion point and control. We also looked at the influence of the interaction between administration time and condition on the rating function parameters.

For the threshold parameter μ, representing the pump number where the probability of an explosion estimated to be .5, there was a significant main effect of

explosion condition ($F(2, 87) = 10.27$, $p < .001$, $\eta^2_G = .191$), administration time ($F(1, 87) = 9.28$, $p = 0.003$, $\eta^2_G = .082$) and interaction between condition and administration time ($F(2, 87) = 8.51$, $p < .001$, $\eta^2_G = .15$) As Figure 6 shows, the pre-task probability ratings show a large difference in perceived probabilities in line with the initial explosion point. There, participants in the high group estimated the balloon to explode with 50% at $\mu = 65.3$ pumps on average, whereas the low group estimated 50% at $\mu = 28.4$ pumps on average, and the control group at $\mu = 45.67$. The significant interaction reflected that with experience, the ratings in the high group decreased, whereas they increased in the low group and the control group. However, only the change in the low group was significant ($t(87) = 4.8$, $p < .001$). For the post-task ratings, the differences between the explosion condition groups were smaller. Although none of the differences remained significant (high vs. control: $t(87) = 1.41$. $p = .15$; low vs. control: $t(87) = 0.47$. $p = .63$; low vs. high: $t(87) = 1.89$. $p = .06$), the rank differences in the mean μ parameter prevail.

For the θ parameter, which represents the slope of the function or how quickly participants switched from low to high probabilities, neither condition ($F(2, 87) = 2.21$, $p = .116$, $\eta^2_G = .048$) nor administration time ($F(1, 87) = 0.05$, $p = .824$, $\eta^2_G = .000$) showed a significant main effect. However, the interaction between condition and administration time ($F(2, 87) = 6.21$, $p < .01$, $\eta^2_G = .125$) showed a significant main effect. This interaction was driven by the significant decrease in the high condition ($t(87) = -2.241$, $p = 0.028$) and significant increase in the low condition ($t(87) = -2.623$, $p = 0.010$). The interaction led to none of the conditions being significantly different at the post-task ratings, which shows how further experience in the task diminished the group differences that were due to the first explosion of the task.

For the γ parameter, which represents the height of the probability ratings in the early stages of the balloon, neither condition, administration time, nor the interaction showed a significant effect and there was no meaningful contrast in differences among condition nor change over time.

To validate the connection between probability rating and behavior, we calculated

Table 2

Correlations between rating function parameters and pumps Study 2

	M	SD	μ_{pre}	θ_{pre}	γ_{pre}	μ_{post}	θ_{post}	γ_{post}	adj. BART
Threshold μ_{pre}	46.44	27.76		0.43***	0.34**	0.53***	0.08	0.18	0.34**
Sensitivity θ_{pre}	20.45	21.34			-0.03	0.38***	0.25*	0.24*	0.33**
Height γ_{pre}	0.05	0.11				0.18	-0.03	0.17	-0.02
Threshold μ_{post}	54.05	24.56					0.28**	0.37***	0.32**
Sensitivity θ_{post}	20.92	9.09						-0.38***	0.06
Height γ_{post}	0.04	0.11							0.29**
adj. BART	38.72	14.66							

Note. p < .001 ***, p < .01 **, p < .05 *. Suffix *pre* refers to the early probability perception-rating task right after the first pump and suffix *post* refers to the late rating task after the thirtieth trial of the BART.

pairwise correlations between each rating function parameter and the adjusted BART score. Table 2 shows the correlation among the parameters as well as with the adjusted BART score for each of the administration times and parameters separately. As can be seen, the threshold parameter μ was generally correlated with the adjusted BART score for both administration times. Significant correlations can also be found for the θ parameter of the first administration and the γ parameter of the second administration. Thus, the threshold parameter μ was generally the best predictor for behavior. When considering the correlations for each parameter, administration and within a condition, and the adjusted BART score, μ was significantly correlated in condition 3 (high condition) in the second administration ($r = .50$, 95% CI [.17, .73], $p = .005$) but in no other parameter values and condition combination. This result could be attributed to the confound of explosion point, which influenced the participant's belief about the balloon in the low and control condition and overruled their private belief. This confound was less strong in the high condition, because only a fraction of participants actually experienced the explosion in the high condition and thus relied more on their private belief.

Mediation Analysis. Finally, we conducted a mediator analysis to investigate whether the difference in pumping behavior between groups was due to differences in subjective probabilities. In the analysis, we used the individual μ parameter of the first rating as a mediator (indirect effect) and the condition (treated as ordinal variable) as a direct effect. The direct effect is the effect of the first explosion point on average pumping in the BART. An indirect effect would then be the effect of the explosion point on perception, which in turn influences the pumping behavior. Thus, a significant indirect effect would indicate that the explosion point indirectly influenced the pumping through the perception of the probabilities. Following first the three-step procedure of Baron and Kenny (1986), a three-step linear regression analysis confirmed that explosion condition was a significant predictor of the μ-parameter, ($b = 9.79$, $SE = 3.41$, $p = 0.004$), and the perception of probabilities (μ) was a significant predictor of the adjusted BART score, $b = .0.18$, $SE = .055$, $p = .001$. Finally, when both variables are considered, the explosion point was no longer a significant predictor of the adjusted BART score ($b = -.17$, $SE = .1.85$, $p = .92$). This would confirm the hypothesis that the explosion point indirectly influenced pumping behavior through the perception of probabilities (an indirect effect). To calculate the effect size P_M, we used bootstrap procedures suggested by (Preacher & Kelley, 2011), resulting in a effect size $P_M = 0.81$ for the indirect effect, which was significant with bootstrapped $CIs = [0.04, 2.45]$.

General Discussion

In two studies, we investigated how risk perceptions influence behavior in a sequential risk-taking task. To probe risk perceptions, we directly asked participants to rate the probability of negative outcomes (i.e., the explosion of a balloon in the BART). Our results showed that these probability ratings did not correspond to the actual probabilities in the BART, nor did they correspond to the assumptions of the most successful cognitive models developed to date. However, we demonstrated that these risk perceptions have a great influence on individual differences of risk-taking behavior in the BART. Finally, our results suggest that risk perceptions are prone to early

experiences in the task, which then orchestrate differences in risk-taking behavior throughout the task. Our study is in line with previous findings that report the importance of early experience in the BART (Koscielniak et al., 2016; Walasek et al., 2014) and even a wider range of tasks (Shteingart et al., 2013). Yet, our results go beyond these findings in that they isolate these effects to participants' risk perceptions. There are two key implications of our results, which we discuss next.

Risk perception and risk taking

Our results clearly indicate the causal role risk perceptions play in risk taking. The approach of using self-reports (of risk perception) is a different approach to that of using cognitive models to investigate the cognitive underpinnings of behavior in the BART (Pleskac, 2008; Pleskac & Wershbale, 2014; Rolison et al., 2012; Wallsten et al., 2005). In fact, our results do not support the assumptions implied by the most successful cognitive model of the BART, which suggests (based on choice data) that the probability of an explosion is treated as small and constant across pumps.

Future studies may systematically test cognitive models that incorporate such probability ratings. Even though our studies were not designed for this goal, as a first attempt of doing so and despite perception ratings were only collected once (Study 1) or twice (Study 2), we ran a thorough modeling analysis (see supplemental material for details of methods), which we briefly summarize here.In total we tested five cognitive models. Four of the models are different variants of the BSR models from Wallsten et al. (2005) and Pleskac (2008), which are a factorial combination of how the probability of an explosion is represented and whether the models assumes there is learning or not. Specifically, two of the models assumed a stationary representation of probabilities in the BART and the other two assumed participants had a correct representation of the BART (the probability of an explosion increased hyperbolically with pump number). Half the models included a Bayesian learning process where participants learned about the probability (stationary) or the number of possible pumps (increasing) with a Bayesian process, or not. In general, models with learning performed better than those

without, and the stationary representation models performed better than non-stationary representation models. These findings are in line with previous applications of these models (Wallsten et al., 2005).

The fifth model that we developed integrated the probability ratings of the perception task into the BSR structure. These models did not include a learning process because with our design, a clear learning process was difficult to assume with only one (Study 1) or two (Study 2) probability rating point(s). This turned out to be a substantial restriction in the model fitting procedure, and BSR models with a Bayesian learning process outperformed the newly implemented perception model in our data. However, our new perception model did perform better than the BSR models that did not incorporate a learning process. This results suggest that probability perceptions of participants account for more behavioral differences than the probabilities that were assumed by the BSR. However, as is in line with our Study 2, learning alters the probability representation in ways that we cannot clearly investigate with our set up. Further studies should address this issue and design experiments that allow for the investigation of probability estimates of participants throughout the task. Additionally, a successful new model should include the influence of the initial experience of the explosion point to accurately account for this effect. A similar model including a so-called *reset* structure has been suggested by Shteingart et al. (2013) for sequential sampling tasks; this potentially could be adapted for a new BART model.

This joint effect of learning and risk perceptions on risk taking raises the question of how important it is to conceal the stochastic structure of the BART as a means to measure risk taking. In fact, several results suggest that when the stochastic structure is more transparent, risk taking becomes much more stable in the task and less reliant on early risk perceptions. For instance, Frey et al. (2015) made the stochastic structure in a similar sequential task more apparent by manipulating the feedback participants received. With "partial feedback" (as provided in the BART), participants did not manage to overcome their early impressions of risk across trials, which systematically shaped risk taking (i.e., relative strong risk aversion as in the BART). Only when

participants were given feedback about the entire stochastic process (in terms of the BART: when the balloon would have exploded), risk taking became less reliant on early risk perceptions and performance more optimal. Similarly, Pleskac (2008) showed that when participants were fully informed about the stochastic structure of the task from the outset (i.e., the probability of a failure at each pump and how it increased), the predictive performance of the task increased compared to when the stochastic structure was hidden. Thus, although obscuring the stochastic structure of the task like the BART makes it possible to understand the cognitive processes involved in risk taking such as the link between say risk-perceptions and risk taking, it also has negative implications in terms of measuring someone's overall risk proclivity. We take up the topic of identifying risk takers next.

Identifying Risk-Takers

The results of the first study are not in line with the original study that we aimed to replicate (Lejuez, Aklin, Jones, et al., 2003). Concretely, we did not find that the adjusted BART score was different for smokers compared to non-smokers. One reason could be minor differences between our study and the one of Lejuez, Aklin, Jones, et al. (2003) such as cultural differences, the difference in sample size (34 non-smokers and 26 smokers instead of 50 of each) or the difference in demographics (the original study only included undergraduate students who were younger than those in our sample and our participants were drawn from a participant pool that includes a variety of different professions and were slightly older). Other minor differences could be that smoking was perceived differently at the time of the original study, more than 13 years earlier. However, we believe that if the BART were a stable predictor for smoking status or unhealthy behavior in general, these minor differences should not have been an issue.

A second possibility is that there is very little relationship between risk preferences such as the adjusted BART score and smoking behavior per se (at least at this point in time). In a review, Harrison, Hofmeyr, Ross, and Swarthout (2015) report that the number of studies that do not find an association between risk preference and

smoking is at least as large as those that do. In addition, studies reporting an association between smoking and risk preferences generally report unstable and marginally significant results. In line with this, many recent studies report no association between the BART and other risk-taking behaviors in the wild (Ba, Zhang, Peng, Salvendy, & Crundall, 2016; Coffey, Schumacher, Baschnagel, Hawk, & Holloman, 2011; Le Bas, Hughes, & Stout, 2015; van Ravenzwaaij, Dutilh, & Wagenmakers, 2011).

Our results highlight a third possibility, namely, that risk perceptions are a key driving factor of actual risk-taking behavior. In particular, in the context of the BART our results demonstrate that the plasticity of risk perceptions — changing in response to a single early explosion point — can have a substantial influence on risk preferences (i.e., the adjusted BART score) that are observed in the task. We should note that in Study 1, we did control the explosion points between participants with half the participants completing the task with the explosion points used in Lejuez et al., and the other half using the inverse of those points. Thus, the impact of the differences in early experiences should have been somewhat minimized in this facet. Nevertheless, it seems possible based on the plasticity of people's risk perceptions in the BART that other factors could differentially influence the perception of risk in the BART. For example, it is well known that the expectations of the experimenter can influence the initial experience of the task for participants and subsequentially bias their behavior (Rosenthal, 1966; Rosnow & Rosenthal, 1997). In the case of smoking, based on our own laboratory experience, it is sometimes difficult to not know when someone is a smoker or not. Considering our findings, other studies with tasks that involve risk taking, especially ones where risk perceptions are a determining factor, should be careful in how they design their studies and how they treat participants, especially if the criterion of interest is easy to identify (e.g., whether one smokes). One solution then is to use blind or double-blind procedures. Another possibility is to consider whether risk perceptions are the key factor of interest. In this case, an alternative solution may be to use a behavioral task that is not so dependent on the risk perceptions, one where the stochastic structure is more transparent, as in the Angling Risk Task (Pleskac,

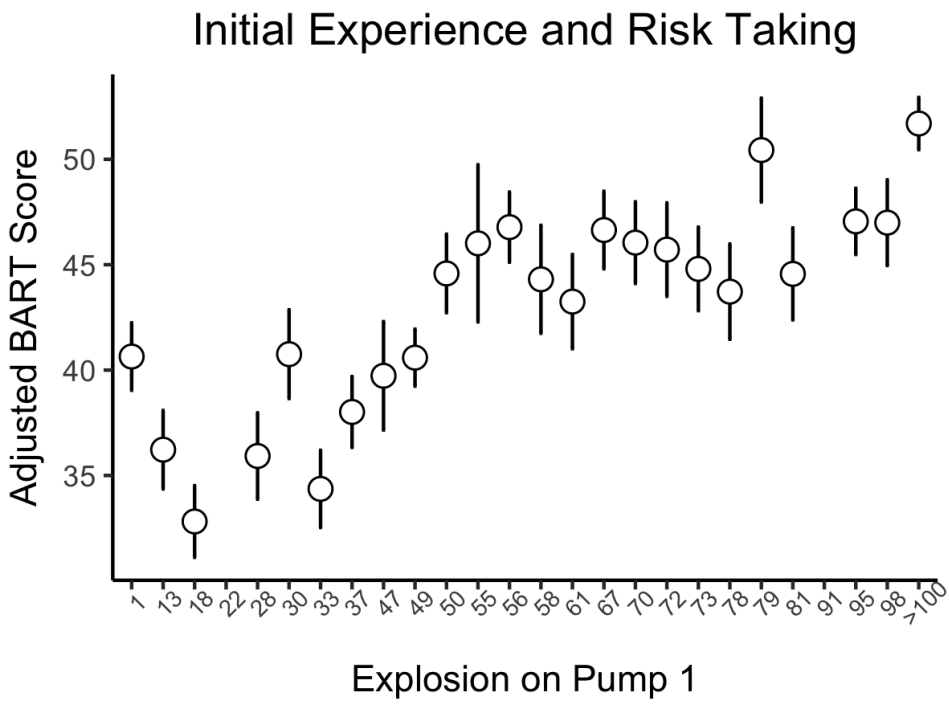

Figure 7. Effect of the initial explosion point in trial 1 on the adjusted BART score, conducted in the Basel-Berlin Risk Study. Of all 1'507 participants of the study, 50% (n = 765) experienced an explosion on the first pump and are included in this analysis. Values above 100 are truncated since few people pumped as far to experience an explosion. Points represent the mean adjusted BART score for all participants that experienced an explosion in the first trial. Error bars represent 95% Confidence Intervals. The correlation between the initial explosion point and the resulting adjusted BART score was significant with $r = .40$, 95% CI [.33, .45], $p < .001$.

.

2008) or the Columbia Card Task (Figner et al., 2009).

Our results may also speak to reports of low consistency between behavioral measures of risk taking and a gap between the latter and self-report measures (Frey et al., in press; Pedroni et al., in press). In particular, in a study based on 39 risk-taking measures and 1,507 participants, behavioral measures such as the BART were found to show poor convergent validity (with other behavioral measures, as well as with self-report measures). Our results suggest that this poor consistency across behavioral measures may result from particular task designs that affected early perceptions of risk.

At least in the case of the BART, data of Frey et al. (in press) support this idea: There, the explosion points were randomized between participants, and as Figure 7 shows the adjusted BART scores were clearly influenced by participants' experience in the first trial. That is, when measuring risk preferences with behavioral tasks, one has to keep in mind that particular task characteristics may influence early risk perceptions, which in turn influence risk-taking behavior in the task. This could be a key driver for the inconsistencies between revealed preferences (obtained from different behavioral tasks), as well as for the gap between revealed and stated preferences (as obtained through self-report measures). One solution to this issue would be to fix the explosion points across participants. Potentially a more effective solution is—to re-emphasize an earlier point—to use tasks where the stochastic structure is much more apparent and thus limiting the potential impact of differences in risk perceptions and learning (see for example Figner et al., 2009; Frey et al., 2015; Pleskac, 2008; Slovic, 1966).

Conclusion

To conclude, our results show that while risk perceptions in the BART do not cohere very well to the actual structure of the task, they do play an important role in the risks people take in the BART. This supports the important role ascribed to risk perceptions in determining actual risk taking. At the same time, our results also demonstrate the critical role early experiences have on shaping risk perceptions, which in turn impact people's overall risk-taking. Taken together our results identify the importance of controlling differences in risk perceptions when using behavioral tasks like the BART to measure risk taking and more broadly the important role early impressions can have on the risks people take.

References

Aklin, W. M., Lejuez, C., Zvolensky, M. J., Kahler, C. W., & Gwadz, M. (2005). Evaluation of behavioral measures of risk taking propensity with inner city adolescents. *Behaviour Research and Therapy*, *43*(2), 215–228. doi: 10.1016/j.brat.2003.12.007

Ba, Y., Zhang, W., Peng, Q., Salvendy, G., & Crundall, D. (2016, January). Risk-taking on the road and in the mind: behavioural and neural patterns of decision making between risky and safe drivers. *Ergonomics*, *59*(1), 27–38. doi: 10.1080/00140139.2015.1056236

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173.

Barratt, E. (1965). Factor analysis of some psychometric measures of impulsiveness and anxiety. *Psychological Reports*, *16*(2), 547 –554.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994, April). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1–3), 7–15. doi: 10.1016/0010-0277(94)90018-3

Coffey, S. F., Schumacher, J. A., Baschnagel, J. S., Hawk, L. W., & Holloman, G. (2011). Impulsivity and risk-taking in borderline personality disorder with and without substance use disorders. *Personality Disorders: Theory, Research, and Treatment*, *2*(2), 128.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. doi: 10.1037/h0047358

Figner, B., Mackinlay, R. J., Wilkening, F., & Weber, E. U. (2009). Affective and deliberative processes in risky choice: age differences in risk taking in the Columbia Card Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 709.

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (in press). Risk preference

shares the structure of major psychological traits. *Science Advances*.

Frey, R., Rieskamp, J., & Hertwig, R. (2015). Sell in May and go away? Learning and risk taking in nonmonotonic decision problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 193–208. doi: 10.1037/a0038118

Gigerenzer, G. (2006). Out of the frying pan into the fire: Behavioral reactions to terrorist attacks. *Risk Analysis*, *26*(2), 347–351. doi: 10.1111/j.1539-6924.2006.00753.x

Harrison, G. W., Hofmeyr, A., Ross, D., & Swarthout, J. T. (2015). Risk preferences, time preferences and smoking behaviour. In *Third Workshop on Behavioural and Experimental Health Economics.*

Helfinstein, S. M., Schonberg, T., Congdon, E., Karlsgodt, K. H., Mumford, J. A., Sabb, F. W., ... Poldrack, R. A. (2014, February). Predicting risky choices from brain activity patterns. *Proceedings of the National Academy of Sciences*, *111*(7), 2470–2475. doi: 10.1073/pnas.1321728111

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523. doi: 10.1016/j.tics.2009.09.004

Hopko, D. R., Lejuez, C. W., Daughters, S. B., Aklin, W. M., Osborne, A., Simmons, B. L., & Strong, D. R. (2006). Construct validity of the balloon analogue risk task (BART): Relationship with MDMA use by inner-city drug users in residential treatment. *Journal of Psychopathology and Behavioral Assessment*, *28*(2), 95–101. doi: 10.1007/s10862-006-7487-5

Jackson, D. N., Hourany, L., & Vidmar, N. J. (1972). A four-dimensional interpretation of risk taking1. *Journal of personality*, *40*(3), 483–501.

Klos, A., Weber, E. U., & Weber, M. (2005, December). Investment Decisions and Time Horizon: Risk Perception and Risk Behavior in Repeated Gambles. *Management Science*, *51*(12), 1777–1790. doi: 10.1287/mnsc.1050.0429

Koscielniak, M., Rydzewska, K., & Sedek, G. (2016). Effects of age and initial risk perception on Balloon Analog Risk Task: The mediating role of processing speed

and need for cognitive closure. *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.00659

Lauriola, M., Panno, A., Levin, I. P., & Lejuez, C. W. (2013, January). Individual differences in risky decision making: A meta-analysis of sensation seeking and impulsivity with the Balloon Analogue Risk Task: Personality and risky decision making. *Journal of Behavioral Decision Making*, *27*(1), 20–36. doi: 10.1002/bdm.1784

Le Bas, G. A., Hughes, M. A., & Stout, J. C. (2015). Utility of self-report and performance-based measures of risk for predicting driving behavior in young people. *Personality and Individual Differences*, *86*, 184–188. doi: 10.1016/j.paid.2015.05.034

Lejuez, C. W., Aklin, W., Daughters, S., Zvolensky, M., Kahler, C., & Gwadz, M. (2007). Reliability and validity of the youth version of the Balloon Analogue Risk Task (BART–Y) in the assessment of risk-taking behavior among inner-city adolescents. *Journal of Clinical Child & Adolescent Psychology*, *36*(1), 106–111. doi: 10.1080/15374410709336573

Lejuez, C. W., Aklin, W. M., Jones, H. A., Richards, J. B., Strong, D. R., Kahler, C. W., & Read, J. P. (2003). The balloon analogue risk task (BART) differentiates smokers and nonsmokers. *Experimental and Clinical Psychopharmacology*, *11*(1), 26–33. doi: 10.1037/1064-1297.11.1.26

Lejuez, C. W., Aklin, W. M., Zvolensky, M. J., & Pedulla, C. M. (2003). Evaluation of the Balloon Analogue Risk Task (BART) as a predictor of adolescent real-world risk-taking behaviours. *Journal of Adolescence*, *26*(4), 475–479. doi: 10.1016/S0140-1971(03)00036-8

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., ... Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84. doi: 10.1037//1076-898X.8.2.75

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal*

of personality and social psychology, *46*(3), 598.

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (in press). The risk elicitation puzzle. *Nature Human Behavior*.

Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 167–185. doi: 10.1037/0278-7393.34.1.167

Pleskac, T. J., Wallsten, T. S., Wang, P., & Lejuez, C. W. (2008). Development of an automatic response mode to improve the clinical utility of sequential risk-taking tasks. *Experimental and clinical psychopharmacology*, *16*(6), 555.

Pleskac, T. J., & Wershbale, A. (2014). Making assessments while taking repeated risks: A pattern of multiple response pathways. *Journal of Experimental Psychology: General*, *143*(1), 142–162. doi: 10.1037/a0031106

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*(2), 93–115. doi: 10.1037/a0022658

Rao, H., Korczykowski, M., Pluta, J., Hoang, A., & Detre, J. A. (2008, August). Neural correlates of voluntary and involuntary risk taking in the human brain: An fMRI Study of the Balloon Analog Risk Task (BART). *NeuroImage*, *42*(2), 902–910. doi: 10.1016/j.neuroimage.2008.05.046

Rogers, R. D., Everitt, B. J., Baldacchino, A., Blackshaw, A. J., Swainson, R., Wynne, K., ... Robbins, T. W. (1999, April). Dissociable deficits in the decision-making cognition of chronic amphetamine abusers, opiate abusers, patients with focal damage to prefrontal cortex, and tryptophan-depleted normal volunteers: Evidence for monoaminergic mechanisms. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *20*(4), 322–339. doi: 10.1016/S0893-133X(98)00091-8

Rolison, J. J., Hanoch, Y., & Wood, S. (2012). Risky decision making in younger and older adults: The role of learning. *Psychology and Aging*, *27*(1), 129–140. doi: 10.1037/a0024689

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century-Crofts.

Rosnow, R., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research.* WH Freeman.

Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption–II. *Addiction (Abingdon, England)*, *88*(6), 791–804.

Schonberg, T., Fox, C. R., Mumford, J. A., Congdon, E., Trepel, C., & Poldrack, R. A. (2012). Decreasing ventromedial prefrontal cortex activity during sequential risk-taking: An fMRI investigation of the Balloon Analog Risk Task. *Frontiers in Neuroscience*, *6*. doi: 10.3389/fnins.2012.00080

Shteingart, H., Neiman, T., & Loewenstein, Y. (2013). The role of first impression in operant learning. *Journal of Experimental Psychology: General*, *142*(2), 476–488. doi: 10.1037/a0029550

Siegrist, M., Keller, C., & Kiers, H. A. L. (2005, February). A new look at the psychometric paradigm of perception of hazards. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, *25*(1), 211–222. doi: 10.1111/j.0272-4332.2005.00580.x

Skeel, R. L., Pilarski, C., Pytlak, K., & Neudecker, J. (2008). Personality and performance-based measures in the prediction of alcohol use. *Psychology of Addictive Behaviors*, *22*(3), 402–409. doi: 10.1037/0893-164X.22.3.402

Slovic, P. (1964). Assessment of risk taking behavior. *Psychological Bulletin*, *61*(3), 220. doi: 10.1037/h0043608

Slovic, P. (1966). Risk-Taking in Children: Age and Sex Differences. *Child Development*, *37*(1), 169–176. doi: 10.2307/1126437

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2011). Cognitive model decomposition of the BART: Assessment and application. *Journal of Mathematical Psychology*, *55*(1), 94–105. doi: 10.1016/j.jmp.2010.08.010

Viscusi, W. K. (1990). Do smokers underestimate risks? *Journal of Political Economy*, *98*(6), 1253–1269.

Walasek, L., Wright, R. J., & Rakow, T. (2014). Ownership status and the representation of assets of uncertain value: The Balloon Endowment Risk Task (BERT). *Journal of Behavioral Decision Making*, *27*(5), 419–432. doi: 10.1002/bdm.1819

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk–taking task. *Psychological Review*, *112*(4), 862–880. doi: 10.1037/0033-295X.112.4.862

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*(4), 263–290. doi: 10.1002/bdm.414

Weber, E. U., & Hsee, C. (1998). Cross-cultural differences in risk perception, but cross-cultural similarities in attitudes towards perceived risk. *Management Science*, *44*(9), 1205–1217. doi: 10.1287/mnsc.44.9.1205

Wichary, S., Pachur, T., Koscielniak, M., Rydzewska, K., & Sedek, G. (2017). Response commentary on effects of age and initial risk perception on balloon analog risk task: The mediating role of processing speed and need for cognitive closure. *Frontiers in Psychology*, *8*. doi: 10.3389/fpsyg.2017.00541

Zuckerman, M. (2007). Sensation seeking and risk. In *Sensation seeking and risky behavior.* (pp. 51–72). Washington: American Psychological Association.

Supplemental Material for: Risk perceptions: Their impact on risk taking and the effect of early experience

Oliver Schürmann[*], Timothy J. Pleskac[**] and Renato Frey[*, **]

[*]University of Basel
[**]Max Planck Institute for Human Development, Berlin, Germany

September 22, 2017

Contents

1 Explosion Points Table

Table 1 shows the explosion points (the number of possible pumps for each trial) of the two studies for each of the conditions. Note that the control condition explosion points are taken from the original study of Lejuez et al. (2003). The reversed order was the result of subtracting the control explosion points from the maximum number of pumps in any trial, which was 128. The explosion points in Study 2 were generated by only changing the first trial's explosion to low (i.e., 13) or high (i.e., 98). All the other numbers in these conditions were kept the same as the original version.

Table 1: List of Explosion Points

	Study 1		Study 2		
Trial	control	reversed	control	low	high
1	65	63	65	13	98
2	104	24	104	104	104
3	39	89	39	39	39
4	80	48	80	80	80
5	21	107	21	21	21
6	8	120	8	8	8
7	121	7	121	121	121
8	96	32	96	96	96
9	60	68	60	60	60
10	38	90	38	38	38
11	64	64	64	64	64
12	101	27	101	101	101
13	26	102	26	26	26
14	34	94	34	34	34
15	41	87	41	41	41
16	121	7	121	121	121
17	62	66	62	62	62
18	95	33	95	95	95
19	75	53	75	75	75
20	13	115	13	13	13
21	70	58	70	70	70
22	112	16	112	112	112
23	30	98	30	30	30
24	88	40	88	88	88
25	9	119	9	9	9
26	72	56	72	72	72
27	91	37	91	91	91
28	17	111	17	17	17
29	115	13	115	115	115
30	52	76	52	52	52

2 Supplemental Descriptive Results

The next section reports descriptive results as well as correlation tables among measures in the two studies. Correlation tables include pairwise rank-order correlations (Spearman's rho).

2.1 Study 1

Tables 2 and 3 list descriptive results and correlations among the tested measures in Study 1. We also ran ANOVAs to measure the predictability of smoking status by the personality questionnaires. Smokers and non-smokers did differ in the sensation-seeking score ($F(1, 98) = 13.58$, $MSE = 24.39$, $p < .001$) and the impulsivity score ($F(1, 98) = 7.49$, $MSE = 77.53$, $p = .007$). This was consistent with the findings of the original BART study.

Table 2: Means and standard deviations of measures in Study 1, separately for smokers and non-smokers

	Smokers		Non-Smokers	
	mean	sd	mean	sd
Adjusted BART	36.84	13.68	38.20	11.93
SS	65.88	4.35	62.24	5.46
IMP	65.92	9.01	61.10	8.59
FTND	6.62	2.95	0.18	1.27
AUDIT	8.30	3.83	5.82	3.31
Lottery	55.96	50.97	35.10	41.21
Benefit	3.55	0.90	3.38	0.81

BART = Ballon Analogue Risk Task. **FTND** = Fagerstom Test of Nicotine Depencence. **AUDIT** = Alcohol Use Dependency Test.

Table 3: Correlations between different measures in Study 1

	1	2	3	4	5	6	7	8
1. Adjusted BART		0.09	0.01	0.04	-0.07	0.16	-0.09	0.05
2. Smoker			-0.34***	-0.28**	-0.90***	-0.33***	-0.22*	-0.11
3. Sensation Seeking				0.40***	0.36***	0.26**	0.09	0.27**
4. Impulsivity					0.29**	0.32**	0.05	0.15
5. FTND						0.32**	0.18	0.15
6. AUDIT							0.03	0.01
7. Lottery								0.02
8. Benefit								

Note. $p < .001$ ***, $p < .01$ **, $p < .05$ *. **adjusted BART** = adjusted BART score. **SS** =Sensation Seeking Score. **IMP** = Impulsivity. **Lottery** = CIP in the lottery task. **Benefit** = mean benefit rating in the benefit task.

2.2 Study 2

Table 4 in the supplemental material provide means and standard deviations for all measures. Table 5 in the supplemental material shows correlations among all measured risk-variables.

Table 4: Means and Standard Deviations of Measures in Study 2 separate for explosion condition

	Control		Low		High	
	mean	sd	mean	sd	mean	sd
BART adj. pumps	42.47	11.80	28.02	13.52	45.68	12.43
Sensation seeking	62.10	5.15	62.40	7.29	63.90	5.47
Impulsivity	59.13	11.24	62.87	12.75	62.67	9.75
FTND	1.13	1.94	1.20	2.34	0.90	1.81
AUDIT	6.13	3.30	5.90	3.43	6.73	3.67
Lottery	59.70	36.72	70.70	36.03	63.67	40.13
Benefit	3.40	0.96	3.75	0.90	3.21	0.77

adjusted BART =adjusted BART score **FTND** = Fagerstom Test of Nicotine Depencence. **AUDIT** = Alcohol Use Dependency Test.

Table 5: Correlations between different measures in Study 2

	1	2	3	4	5	6	7
1. Adjusted BART		0.02	-0.10	0.05	0.16	-0.05	-0.19
2. Sensation Seeking			0.25*	0.07	0.28**	-0.05	0.06
3. Impulsivity				-0.06	-0.09	0.12	-0.04
4. FTND					0.07	-0.03	-0.12
5. AUDIT						-0.01	-0.05
6. Lottery							-0.07
7. Benefit							

Note. $p < .001$ ***, $p < .01$ **, $p < .05$ *. **adjusted BART** = adjusted BART score. **FTND** = Fagerstom Test of Nicotine Depencence. **AUDIT** = Alcohol Use Dependency Test. **Lottery** = CIP in the lottery task. **Benefit** = mean benefit rating in the benefit task.

3 Benefits Task

3.1 Methods

Studies investigating subjective probabilities typically also investigate the subjective benefit of different risks people take (Weber et al., 2002; Blais & Weber, 2006). With the BART benefit task, we aim to measure participants' subjective benefits from pumping the balloon. Participants were shown the same 11 sizes of balloons as in the perception rating task. Participants were then asked to indicate their subjective benefit for pumping the balloon on more time, given the current size of the balloon. Participants rated their subjective benefit on a 5 point Likert scale ranging from 1 to 6 for each of the balloon sizes. In both studies, the benefits task was conducted right after the perception task that followed the BART.

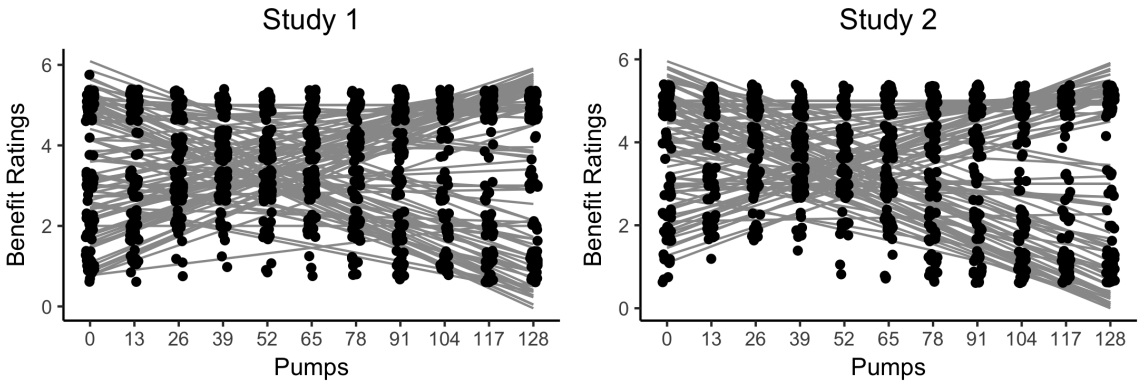

Figure 1: Benefit ratings and linear regression lines for each participant over all 11 ratings. There appeared to be three clusters of people, each of which characterized by a specific tendency of subjective benefits over increasing pumps of the balloon. Some participants rated the benefits as increasing over the size, some as decreasing, and some as stationary.

3.2 Results

3.2.1 Study 1

To analyze the behavior in the benefits task in detail, we fitted a separate linear function over all ratings for each participant. Figure 1 shows the distribution of benefit ratings as well as the fitted linear regression lines for each participant. The slope parameter β of each linear regression gave an indication whether the benefit of pumping increased, decreased or was stationary over increasing balloon sizes for a certain participant (i.e., increasing pumps). For 29% of participants the benefit rating decreased when the balloon got bigger ($\beta < 0$), for 20% of participants the benefit ratings were stationary ($\beta = 0$) and for 51% of participants the benefit rating increased ($\beta > 1$). We dummy-coded participants based on their change of benefit rating over the 11 balloons into three groups; increasing benefit ($\beta > 0$), decreasing benefit ($\beta < 0$) and stationary benefit ($slope = 0$) to test whether the groups differed in pumping behavior in the BART or smoking status.

In summary, we did not find any relationship between the benefit ratings and risk-taking behavior in the BART. There was no correlation between mean benefit ratings (aggregated over all 11 ratings) and the adjusted BART score ($r = .11$, 95% CI [−.09, .30], $p = .293$). The slope β of the benefit function did not have any correlation with behavior in the BART either ($r = .11$, 95% CI [−.09, .30], $p = .293$) and an ANOVA revealed no difference in adjusted BART score between the benefit groups ($F(2, 97) = 0.22$, $p = .807$, $\eta^2_G = .004$). Finally, we did not find any difference of mean benefit ratings for smokers and non-smokers ($F(1, 98) = 1.04$, $p = .311$, $\eta^2_G = .003$). Moreover, as Figure 1 shows there was hardly any systematicity in the use of the scale therefore we chose not to use the scale in the analysis.

3.2.2 Study 2

Out of curiocity we had participants complete the benefits task again in Study 2. Results in the benefit task of Study 2 were similar to the ones of Study 1. The right panel of Figure 1 shows the benefit ratings for each of the balloon sizes and the linear regression lines, fitted to each participants' ratings. As in Study 1, we observe the three tendencies in benefit rating change with increasing

pumps. For 46.7% of participants, the benefit rating decreased with balloon size ($\beta < 0$), for 18.9% of participants it was stationary ($\beta = 0$) and for 34.4% of participants, the benefit rating increased ($\beta > 1$). We dummy-coded participants based on their change of benefit rating again over the 11 balloons into three groups; increasing benefit ($\beta > 0$), decreasing benefit ($\beta < 0$) and stationary benefit ($slope = 0$).

In Study 2 we did find a relationship between the benefit ratings and risk-taking behavior in the BART. Despite there was no significant correlation between mean benefit ratings (aggregated over all 11 ratings) and the adjusted BART score ($r = -.19$, 95% CI $[-.38, .02]$, $p = .077$), the slope β of the benefit function correlated with the adjusted BART score ($r = -.32$, 95% CI $[-.50, -.12]$, $p = .002$) and an ANOVA revealed a difference in adjusted BART score between the benefit groups ($F(2, 87) = 3.63$, $p = .031$, $\eta^2_G = .077$). However, the effect was insignificant when we controlled for the influence of the condition (benefit group: $F(2, 81) = 2.09$, $p = .131$,
$eta^2_G = .049$, condition: $F(2, 81) = 13.63$, $p < .001$,
$eta^2_G = .252$).

3.2.3 Discussion and Conclusion

Given the inconsistency of subjective benefit ratings in both studies and the apparent inexistent (or barely, if any) connection to behavior in the BART, we do not interpret the benefit rating task in detail. The most trivial explanation for the results is that participants were unsure about the exact question they were to answer. This may stem from the fact that in the BART, the benefit of a single pump is specified (i.e., 1 cent for each pump) and known to the participant. Thus, asking something obvious could have lead to some confusion about the true meaning of the question and triggered the inconsistent behavior. Another possibility is that the differences might reflect the different framings participants take with regard to the task with some including an aversion to a balloon exploding, others consider the future potential earnings, while others are considering just the next potential earning. Regardless, the lack of relationship between how people rated the benefits and their actual behavior implies this particular rating gives us little insight into their choice to pump or not.

4 Lottery Task

4.1 Study 1

We investigated how choice behavior in the BART deviated from choices between monetary lotteries that corresponded with the pump options in the BART. The goal of this task was to examine a possible description-experience gap with a task like the BART (Hertwig & Erev, 2009). For instance, at pump number 10, participants could be thought of as choosing between €0.50 for sure or a chance to earn €0.55 with a probability of $\frac{118}{119}$ otherwise nothing. To examine the correspondence between the choices and the BART and choices between equivalent monetary lotteries, we estimate the equivalent pump i where participants were indifferent between the sure gain and the lottery. The pump representing this choice indifference point (CIP) is equivalent to the point where participants stop pumping the balloon, and can thus be directly compared to the adjusted BART score. We tested how the lottery CIP and the adjusted BART score differed and quickly discuss the results.

4.1.1 Methods

The lotteries represented the exact choice-options in the BART. That is, the decision on each pump opportunity is between cashing in and earning a certain amount for sure or pumping with the chance

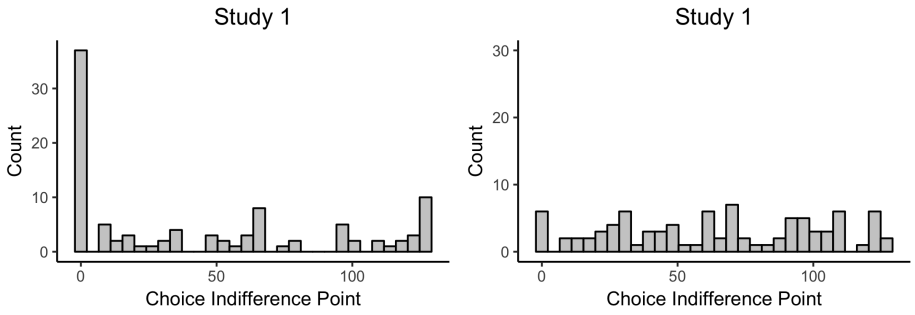

Figure 2: Mean CIP for each participant, based on the choices in the lottery task for both studies. **Left Panel:** Results from Study 1, **Right Panel:** Results from Study 2

of earning an additional €0.05 but risk losing it at all. To better align with this decision we asked participants to make a series of choices between a lottery and a sure thing that corresponded to a particular pump i in the BART. The options were described in the form shown below.

	Earn €0.05
Option A	with the probability $\frac{128-i}{128-i+1}$; otherwise get 0
Option B	Earn $i\times$ €0.05 for sure

To estimate the CIP in the lottery task, we used a psychophysical up-down method instead of presenting every lottery for all balloon sizes (i.e., size 1 up to size 128). The up-down method works as such, as that the option space (i.e., all pump opportunities i equaling 1 to 128) is systematically tested by asking the extreme points first (i.e., 1 and 128), then the midpoint (i.e., $i = 64$). Then, the lottery "jumps" up if the participant chooses the lottery or down if the participant chooses the safe option. The jump size n starts at 128 and is then halved after every jump (first, $[n_1 = 128, n_2 = 64, n_3 = 32, ..., n_i = 1]$). Thus the participant first gets presented the gamble in the first pump (i.e., $i = 1$, "get 0 for sure or €0.05 with a probability 99%, otherwise nothing"). If the participant accepted the risky gamble, he was presented the highest possible pump ($i = 128$), which is choosing between option 1:€3.7 for sure or option 2: €0 for sure. The first two gambles are trivial since there is a completely dominated option. Thus participants would choose the risky option for the first gamble $i = 1$ and the safe option for the second gamble ($i = 128$). Then the jump would go down by half the option space left and gamble $i = 64$ is presented. The next jump was halved again and went down if the participant chose the safe option ($i = 32$), or up if the participant chose the risky option ($i = 97$). For jumping through the option space in this manner, only eight gamble questions are needed instead of asking all the way up to the CIP.

4.1.2 Results and Discussion

On average, participants in Study 1 indicated a mean CIP of 43.9 ($SD = 47.46$). The distribution of CIPs was largely skewed as can be seen in Figure 2 left panel. Over 36% of participants indicated to have a CIP of 1, which means they would never have chosen the risky option, which would be the equivalent of not pumping the balloon at all. Finaly, the CIP was not correlated with the adjusted BART score ($r = -.12$, 95% CI [$-.30$, .08], $p = .253$).

These results could be attributed to the differences in task structure in the BART and the lottery task. In the lottery task, each gamble offered two options, the safe option (cash out) or take the

risky option with two outcomes, either nothing or the amount in the bank. In the BART, after choosing the risky option, participants can go on playing. Thus every choice option (i.e., pump opportunity) in the BART offers as option 1 a safe amount, and as option 2 a risky option which included the possibility of winning the amount in the bank and *in addition* the possibility to go on playing the game and win more. Therefore, the risky option (i.e., pumping) is more attractive in the BART than in the lottery task, which explains the lottery task to be an insufficient competitor for descriptive presentation of the lotteries in the BART.

4.2 Study 2

The lottery task in Study 2 was adapted from the one in Study 1 but aimed at resolving some of the possible issues from Study 1. In the lottery task of Study 2, we aimed at better accounting for the fact that the risky option (pumping the balloon), also includes being able to play along. That is, in the BART, the decision on each pump opportunity is either take the safe option, and cash in, or take the risky option and pump more. Pumping more gives the chance of earning an additional €0.05, but also to be able to go on playing. Thus the risky option always also represents the possibility of pumping more and earning even more. To better align with this decision environment, we added the possible expected value that comes from going on playing. Thus we asked participants to make a series of choices between a lottery and a sure thing that corresponded to a particular pump i in the BART.

4.2.1 Methods

The options were described in the form shown below.

	Earn €0.05 plus the opportunity win additional €A on average
Option A	with the probability $\frac{128-i}{128-i+1}$; otherwise get 0
Option B	Earn $i\times$ €0.05 for sure

The average additional earnings was calculated as average expected value of,

$$A = \frac{\sum_{i+1}^{128}(\frac{128-i}{128-i+1} \times 0.05)}{128-(i+1)}.$$

To find the CIP for each participant, we again used the psychophysical up-down method described in Study 1.

4.2.2 Results and Discussion

Distributions of the CIPs in Study 2 can be seen in the right panel of Figure 2. Substantially fewer participants now only chose the safe option. Only 6% of participants always chose the safe option, compared to 36% in Study 1. The CIP from the lottery task was still not correlated with the adjusted BART score ($r = -.04$, 95% CI $[-.25, .16]$, $p = .681$).

The results of the lottery task in Study 2 indicated two things. First, participants chose the risky option more often than in the version of Study 1. We conclude that the addition of the average expected reward to the risky option did make the risky option more attractive. Second, when presented with a descriptive version of gambles in the BART, participants seem to be more risk-seeking than in the task. In fact, a CIP of 64 represents the mathematically optimal strategy for this gamble. So the mean CIP in the lottery task represents risk-neutral behavior. Overall, these

results indicate that when participants know the real probabilities in the task, on average, on average their behavior corresponds with the mathematically dominant strategy. If however, probabilities are unknown to the decision-maker (such as in the BART), participants need to rely on their perceptions and estimate the probability which can result in seemingly risk-averse behavior.

5 Cognitive Modeling

We conducted a thorough mathematical cognitive modeling analysis for the behavior in the BART in both studies based in the models of (Wallsten et al., 2005; Pleskac, 2008). We fitted 5 models in total. 4 different variations of the Bayesian Sequential Risk-Taking Model (BSR) and one model that is derived from the BSR models but incorporates the probability ratings of the perception task. We briefly summarize each model briefly, provide how the model fit the data using Maximum Likelihood estimation and compare the models using the Bayesian Information Criterion (BIC). Finally, we discuss the modeling results in light of the findings of our main analyses.

5.1 Bayesian Sequential Sampling Model

The BSR model assumes that Decision-Makers (DM) use three cognitive processes to make decisions in the BART task: evaluation, response, and learning. In the evaluation stage, the DM defines a target pump that maximizes his value, which he probabilistically pumps towards, defined by the response selection process. Finally, depending on the model variation, the participants learns about the properties in the task with a Bayesian updating process, or has the same target for all trials. Each of the processes are defined by at least one free parameter that describes how each participant individually differs in each of the three processes. The four variations of the BSR we tested differ by two properties; the way subjective probabilities are represented and whether the model incorporates a learning process or not. A detailed formal description of the models can be viewed in Wallsten et al. (2005) and Pleskac (2008). Here we briefly outline the models and provide the formulas for the different steps.

5.1.1 Stationary Representation Model with Learning (BSR st-l)

The model assumes that the DM evaluates the expected gain on trial h for each choice option i is given by

$$v_{v,i} = \mu_h(i) * (ix)^{\gamma^+}, \tag{1}$$

where $\mu_h(i)$ represents the subjective probability and γ is like prospect theory's diminishing sensitivity parameter. In the stationary model, the subjective probability of i successes is assumed to be stationary over all possible pumps and equals $\mu_h(i) = \hat{q_h^i}$. Thus, the target pump G for trial h is then derived by optimizing equation 1, which can be expressed as

$$G_h = \frac{-y^+}{ln(\hat{q}_h)} \tag{2}$$

.

The DM will then probabilistically pump towards this target with probability $r_{h,i}$ to pump on trial h option i given by the following response rule

$$r_{h,i} = \frac{1}{1 + exp(\beta d_{h,i})} \tag{3}$$

with $d_{h,i} = i - G_h$ and β is a free parameter representing how consistently DMs follow their targeted evaluation.

Finally, after every trial, the DM updates his subjective probability $\hat{q}_i$ based on the observations made in the trial using Bayes rule

$$f_{h+1}(q|c_1, d_1, ..., c_n, d_n) = \frac{p(c_1, d_1, ..., c_h|q)f_1(q)}{\int p(c_1, d_1, ..., d_h, d_h|q)f_1(q)dq} \tag{4}$$

with c_h denoting the total number of successes for a particular trial h and d_h whether the balloon exploded on that trial ($d_h = 1$) or not ($d_h = 0$). The learning process is defined individually by two learning parameters $\hat{q}_1$ denoting the mean and ω_1 denoting the variance of a beta distribution that represents the prior belief of the participant.

5.1.2 Nonstationary Representation model with Learning (BSR ns-l)

The nonstationary model is based on the same basic processes, evaluation, response selection and learning, but assumes an increasing subjective probability. However, the nonstationary model "assumes that DMs adopt a correct mental representation but remain uncertain about the precise properties of the task"(Pleskac, 2008). DMs are uncertain of the maximum number of possible trials, n. Thus, the evaluation of the probability of success at pump i is denoted by

$$\frac{(\hat{n}_h - i)}{\hat{n}_h} \tag{5}$$

resulting in the following evaluation formula that defines the target pump G_h on balloon h with the closed form

$$G_h = \frac{\hat{n}\gamma^+}{\gamma^+ + 1}. \tag{6}$$

The response selection stage is the same as in the stationary model, denoted by Equation 3. Finally, the learning rule for the increasing model

$$p_{h+1}(n|c_1, d_1, ..., c_n, d_n) = \frac{(c_1, d_1, ..., c_h|n)p_1(n)}{\sum_n p(c_1, d_1, ..., d_h, d_h|n')p_1(n')} \tag{7}$$

5.1.3 Non-learning Models (BSR_st-nl and BSR_ns-nl)

The two non-learning models are variations of the stationary and nonstationary representation models described above, respectively. The evaluation and response selection stages are the same as the respective learning models. The main exception is, however, that they do not include a learning component. These models assume, suggest that the participant does not update his belief about the explosion probability of the balloon, and with this has the same target number of pumps for every trial. The non-learning models are psychologically implausible. However, they allow to better determine the separate influence of learning and subjective probabilities.

5.2 Perception Model

To better account for and investigate the impact of risk perceptions, we adapted the non-learning model framework of the BSR and included the self-reported probability ratings from the two experiments. As the BSR, this model assumes a target number of pumps for each participant by

maximizing the expected gain of pumping on a certain trial i. The model uses the same estimation of options as the BSR with the exception that it integrates the exact ratings of the participants to determine the probability of success. Thus, the evaluation stage of the model is formulated by,

$$v_{v,i} = pr(i) * (ix)^{\gamma^+}, \tag{8}$$

where $pr(i)$ denotes the probability of a success, given the joint subjective probability ratings of a participant up to pump i. The γ parameter is akin to the other models and represents a weighting of the possible payoffs.

We weighted the probability ratings with a Prelec weighting function, typically used in prospect theory models (Prelec, 1998). The weighting function has one free parameter α that defines how much under- or overweighting of participants' probability ratings is reflected in their behavior. The perception model has thus three free parameters. A γ parameter for weighing the payoffs, a α parameter of the weighting function and a behavioral consistency parameter β of the response model.

For Study 2, the perception model incorporated both the pre- and the post-task probability ratings by linearly increasing the ratings from the pre-task rating to the post-task rating.

5.3 Model Estimation

The models were fit to each participants' choice data for all rounds of the BART using maximum likelihood estimation methods described in Appendix A of Pleskac (2008). The models were compared using the Bayesian Information Criterion (BIC), which has the advantage that it takes into account how parsimonious the model is (gives a penalty for higher numbers of free parameters).

5.4 Results and Discussion

5.4.1 Study 1

Table 6 summarizes the fit (given in mean BIC and mean Log-Likelihood) of each of the five models used for Study 1, how many participants were described best by the respective models and the number of free parameters (df). Consistent with the previous administrations of the models, learning models showed a better fit than non-learning models and the models assuming stationary subjective probability representation fit the data better than those that assume a non-stationary representation. The perception model fit the data better than the BSR non-learning BSR models but less than the BSR learning models.

5.4.2 Study 2

Table 8 summarizes the fit of each model we tested in Study 2. Consistent with the previous administrations of the models and with Study 1, the learning models fit the data better than the non-learning models and the models assuming a stationary subjective probability representation perform better than those that assume a non-stationary representation. Finally, as in Study 2 the perception model fit the data better than the BSR non-learning models, but no than the learning models.

5.4.3 Discussion

In both studies, the perception models did perform better than the non-learning versions of the BSR, which supports our assumptions that the self-reported subjective probabilities were a driving

Table 6: Model Comparisons of Study 1 showing the number of free parameters per model (df), the number of DMs best fit by each model according to BIC, and the mean values of the two fit statistics (BIC and MLL)

model	df	# DMs best fit	Mean.BIC	Mean.MLL
BSR ns-l	4	22	119.22	-73.33
BSR st-l	4	78	114.18	-70.80
BSR ns-nl	2	0	161.18	-87.45
BSR st-nl	2	0	160.72	-87.22
Perception Model	3	0	147.49	-80.60

Note. DM = decision maker; BIC = Bayesian information criterion; MLL = Maximum Log Likelihood; st = stationary; ns = non-stationary; l = learning; nl = non-learning

Table 7: Model Comparisons of Study 1 for non-learning models

model	df	No..of.DMs.best.fit	Mean.BIC	Mean.MLL
BSR ns-nl	2	19	-161.1813	-87.44645
BSR st-nl	2	16	-160.7292	-87.22040
Perception model	2	65	-147.4970	-80.60427

Note. Showing the Number of Free Parameters per Model (df), the Number of DMs Best Fit by Each Model According to BIC, and the Mean Values of the Two Fit Statistics (BIC and MLL). DM = decision maker; BIC = Bayesian information criterion; MLL = Maximum Log Likelihood

Table 8: Model Comparisons of Study 2, Showing the Number of Free Parameters per Model (df), the Number of DMs Best Fit by Each Model According to BIC, and the Mean Values of the Two Fit Statistics (BIC and MLL)

model	df	No..of.DMs.best.fit	Mean BIC	Mean MLL
BSR ns-l	4	34	123.70	-75.55
BSR st-l	4	56	119.15	-73.27
BSR ns-nl	2	0	-186.45	-100.07
BSR st-nl	2	0	-186.45	-100.07
Perception Model	3	0	170.50	-92.10

Note. DM = decision maker; BIC = Bayesian information criterion; MLL = Maximum Log Likelihood; st = stationary; ns = non-stationary; l = learning; nl = non-learning

factor for behavior in the BART. However, the perception models did not outperform the learning models, which is not surprising, since the non-learning models assume participants to have the same target of pumping over all trials and do not adjust their behavior based on their experience with the

Table 9: Model Comparisons of Study 2 for non-learning models

model	df	No..of.DMs.best.fit	Mean.BIC	Mean.MLL
BSR ns-nl	2	24	-186.4553	-100.07
BSR st-nl	2	16	-186.4553	-100.07
Perception Model	3	50	-170.50	-92.10

Note. Showing the Number of Free Parameters per Model (df), the Number of DMs Best Fit by Each Model According to BIC, and the Mean Values of the Two Fit Statistics (BIC and MLL). DM = decision maker; BIC = Bayesian information criterion; MLL = Maximum Log Likelihood

task. Further comparisons and the implementation of a model that integrates the findings of this study should be the scope of future studies.

References

Blais, A.-R., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, *1*(1). Retrieved 2016-10-11TZ, from `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1301089`

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523. doi: 10.1016/j.tics.2009.09.004

Lejuez, C. W., Aklin, W. M., Zvolensky, M. J., & Pedulla, C. M. (2003). Evaluation of the Balloon Analogue Risk Task (BART) as a predictor of adolescent real-world risk-taking behaviours. *Journal of Adolescence*, *26*(4), 475–479. doi: 10.1016/S0140-1971(03)00036-8

Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 167–185. doi: 10.1037/0278-7393.34.1.167

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

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk–taking task. *Psychological Review*, *112*(4), 862–880. doi: 10.1037/0033-295X.112.4.862

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*(4), 263–290. doi: 10.1002/bdm.414

Oliver Schürmann
CV

Basel, 10.10.2017

Oliver Schürmann

Languages	German (native), English (C2, CPE) French (B2)	
Research Interests	- Decision-making under risk - Risk-taking in the wild - Cognitive modeling - Neuronal underpinnings of decision-making - Risky Decisions in the Wild	
Education	Gymnasium Bäumlihof *Major in Biology and Chemistry*	*2000-2005*
	Independence Community College, Kansas, USA *Major in General Studies, Studying abroad Program*	*2005-2006*
	University of Basel *Bachelor of Science in Psychology*	*2007-2010*
	University of Basel *Master of Science in Psychology* *Center for Economic Psychology*	*2010-2012*
Academic Experience	Student Assistant *Center for Economic Psychology, University of Basel*	*Oct. 2010 - Oct. 2013*
	Research Internship, University of Oslo, Norway *With Guido Biele, Department of Psychology, UiO*	*Oct. & Nov. 2012*
	Graduate Student, Center for Economic Psychology, Department of Psychology, University of Basel, Switzerland	*Oct. 2013-Sep. 2017*
	Visiting Researcher, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin	*May 2015-Nov. 2017*
	Postdoctoral Researcher, Center for Economic Psychology, Department of Psychology, University of Basel, Switzerland	*Nov. 2017-Feb. 2018*
	Scientific Advisor, Istituto Marco Ronzani, Basel	*Feb. 2018-present*
	Postdoctoral Researcher, Chair for Cognitive Science, Swiss Federal Institute of Technology (ETH) Zurich	*May 2018-present*
Grants	"The Underlying Decision-Making Processes of Traffic Related Risk-Taking Behavior", Doc.CH Grant of the SNSF (P0BSP1_148884) 224'000 CHF,	*Oct. 2013 – 2016*
	Mobility grant, Research Stay at the Center for Adaptive Rationality, MPI, Berlin, SNF Project P0BSP1_148884 / 2, 4800 CHF	*April 2015*
Courses	SPM Course 2011 Zurich	

Oliver Schürmann
CV

Basel, 10.10.2017

& Workshop

Course on statistical parametrical mapping of fMRI data *January 2011*

Cambridge Certificate of Proficiency in English
Grade B *June 2011*

5th JDM Workshop for early Career Researchers
University of Basel *July 2012*

Winterschool on Learning, Bounded Rationality and Decision-Making,
Technion, Haifa, Israel *January 2014*

7th JDM Workshop for early Career Researchers *July 2014*
University of Mannheim

Cognitive Modeling Summer School, Laufen, Germany *August 2014*

Workshop: Bayesian analysis with BayesFactor, UCI, Irvine, USA *November 2014*

University Based Courses: *Project Management, Data Analysis, Scientific Writing, Writing productivity, Leadership Training, Presentation Skills, Negotiation skills. Conflict Management skills. Certificates for each course available on demand.*

Conference Talks & Posters

Schürmann, O & Klucharev, V. (2012). *DOTS Lying-Task, The Influence of Descriptive Norms on Deception*. Talk at the 5th JDM Workshop for Young Researchers, Basel, Switzerland.

Schürmann, O., Pedroni A. & Rieskamp J. (2014). *The Underlying Decision-Making Processes of Traffic-Related Risk-Taking Behavior.* Talk at the 7th JDM Workshop for Early Career Researchers, Mannheim, Germany.

Schürmann, O., Pedroni A., Frey, R., Hertwig, R., & Rieskamp J. (2014). *Risk-Taking Behavior in Traffic.* Talk at the 6th Bernoulli Workshop in Economics and Psychology. University of Basel, Switzerland.

Schürmann, O., Pedroni A., Frey, R., Hertwig, R., & Rieskamp J. (2014). *The Underlying Decision-Making Processes of Traffic-Related Risk-Taking Behavior.* Poster presented at the 35th Annual Conference of the Society for Judgment and Decision Making, Long Beach, California, USA.

Schürmann, O., Pedroni A., Frey, R., Hertwig, R., & Rieskamp J. (2014). *Risk-Taking Behavior in Traffic – Measuring and Predicting Risk-Taking in a Real-Life Traffic Task*. Talk at the 1st PhD Conference of the PhD Program in Social, Economic, and Decision-Making Psychology, University of Basel, Switzerland. rik

Schürmann, O., Pedroni A., Frey, R., Hertwig, R., & Rieskamp J. (2015). *Measuring and Predicting Risk-Taking in a Real-Life Traffic Task*. Talk as part of the symposium: A Multimethod Approach to Measure Risk-Taking Behavior, Frey R. & Pedroni A. at the 57th Conference of Experimental Psychologists (TeaP), Hildesheim, Germany.

Schürmann, O., Frey, R., Hertwig, R., Riekamp, J., & Pedroni, A. (2015). *Predicting risk-taking behavior in a real-life traffic task.* Poster at the Subjective Probability, Utility, and Decision Making Conference (26th SPUDM). Budapest, Hungary.

Schuermann, O., Pleskac, T. J., Frey, R., & Hertwig, R., (2016). *Risk Perception in the Balloon Analogue Risk Task* 9th JDMx Meeting for Young Researchers, Basel, Switzerland.

Schuermann, O., Pleskac, T. J., Frey, R., & Hertwig, R., (2016). *Risk Perception in the Balloon Analogue Risk Task* Meeting of the Society for Neuroeconomics. Berlin, Germany.

Oliver Schürmann
CV

Basel, 10.10.2017

Schuermann, O., Andraszewicz, A., & Rieskamp, J., (2017). *Reliability and Validity of Economic Risk Elicitation Measures* 10th. JDMx Meeting for Young Researchers, Bonn, Germany

Schuermann, O., Pleskac, T. J., Frey, R., & Hertwig, R., (2017). *Risk Perception in the Balloon Analogue Risk Task*. Talk at the 27th Bienial Subjective Probability, Utility, and Decision Making Conference (SPUDM). Haifa, Israel

Research Skills

Software & Programming Skills
Presentation, e-Prime, Shell, JAVA, FSL, SPM, MatLab, R, SPSS,

Experimental Skills
planning, conducting and analyzing of Behavioral - and fMRI-Experiments