Decision
2021, https://dx.doi.org/10.1037/dec0000127

Through the Window of My Mind: Mapping Information Integration and the Cognitive Representations Underlying Self-Reported Risk Preference

Markus D. Steiner[1], Florian I. Seitz[1], and Renato Frey[1,2]
[1]University of Basel
[2]Princeton University

A person's risk preference may determine significant life outcomes (e.g., in finance or health), and people are therefore routinely asked to report their risk preferences in various scientific and applied contexts. Yet, still little is known concerning the cognitive underpinnings of this judgment-formation process. We ran two studies (N = 250, and N = 150 in a retest) implementing the process-tracing method of *aspect listing*, to investigate the information-integration processes underlying people's self-reports by means of cognitive modeling (RQ1), as well as to examine people's cognitive representations of their risk preferences (RQ2). Our analyses indicate that interindividual differences in self-reported risk preferences can be modeled well based on the listed aspects' properties of evidence and substantially better than using sociodemographic variables as predictors. Specifically, to render self-reports people appear to integrate the *strength of evidence* of multiple aspects sampled from memory. These aspects further revealed that people's cognitive representation of their risk preferences mostly relate to the magnitudes of outcomes and often to explicit trade-offs between positive and negative outcomes, in line with a risk–return perspective. Crucially, within participants the strength of evidence of the listed aspects remained highly stable across the two studies (RQ3), and changes therein were closely related to changes in self-reported risk preference (RQ4). In sum, our findings provide insight into the cognitive processes of how people render self-reports of their risk preferences, suggest an explanation for the well-documented temporal stability thereof, and thus corroborate the internal validity of this measurement approach.

Keywords: risk preference, self-report measures, process tracing, aspect listing

"Are you generally a person who is willing to take risks or do you try to avoid taking risks?" Chances are that a person is confronted with this or a similar question in numerous settings, such as when discussing private investments with a financial advisor (Balatel et al., 2013; Ferrarini & Wymeersch, 2006) or when taking part in one of the many panel studies that are routinely conducted around the world (e.g., the German Socio-Economic Panel, SOEP; Dohmen et al., 2011; Lejarraga, Frey, Schnitzlein, & Hertwig, 2019). These measurement attempts are not surprising, given that people's risk preferences may shape important life outcomes, such as financial bankruptcy as a consequence of risky investments, or addiction as a consequence of experimenting with substance use. But how do people render judgments concerning their own risk preferences? And can such *stated preferences* indeed be considered valid?

In psychology and the behavioral sciences more generally, self-reports have a long-lasting and successful tradition (Cronbach, 1946; Galton, 1874; Guttman, 1944; Likert, 1932; Thurstone, 1927, 1928). For example, self-report measures were instrumental in the discovery of major constructs such as the Big Five personality dimensions (e.g., McCrae & Costa, 1987) and continue to be an important tool for studying concepts such as grit (e.g., Duck-

Markus D. Steiner, Florian I. Seitz, Renato Frey, Center for Cognitive and Decision Sciences, Department of Psychology, University of Basel (all authors), and Behavioral Science for Policy Lab, Andlinger Center for Energy and the Environment, Princeton University (R.F.).

This work was supported by a grant of the Swiss National Science Foundation (PZ00P1_174042)) provided to Renato Frey. We thank Olivia Fischer and Samuel Zeiser for their help with the content-ratings of the aspects, Jana Jarecki for helpful comments on an earlier version of this work, and the members of the Center for Cognitive and Decision Sciences for valuable input. We thank Laura Wiles for proofreading. Data, analysis code, screenshots of the experimental paradigm, and the preregistration are available at https://osf.io/gndjw.

Corresponding author: Markus D. Steiner, Department of Psychology, University of Basel, Missionsstrasse 60/62, 4055 Basel, Switzerland. E-mail: markus.steiner@unibas.ch

worth, Peterson, Matthews, & Kelly, 2007) or well-being (e.g., Diener, 1984; Kahneman & Deaton, 2010). Crucially, self-report measures not only are easy to implement (e.g., Dohmen et al., 2011; Duckworth & Yeager, 2015) but often also exhibit desirable psychometric properties. To illustrate, in the context of risk preference and closely related constructs, self-report measures were found to have high convergent validity, test–retest reliability, and predictive validity (Beauchamp, Cesarini, & Johannesson, 2017; Duckworth & Yeager, 2015; Frey, Pedroni, Mata, Rieskamp, & Hertwig, 2017; Galizzi, Machado, & Miniaci, 2016; Lönnqvist, Verkasalo, Walkowitz, & Wichardt, 2015; Mata, Frey, Richter, Schupp, & Hertwig, 2018; Rohrer, 2017). By contrast, their behavioral counterparts—that is, game-like tasks such as monetary lotteries, which may be indispensable for applications such as examining the functional neural architecture of risk preference (e.g., Tisdall et al., 2020; Tom, Fox, Trepel, & Poldrack, 2007)—generally tend to be more intricate to implement (see Andreoni & Kuhn, 2019; Pedroni et al., 2017) and often fail to meet fundamental measurement properties (e.g., Beauchamp et al., 2017; Berg, Dickhaut, & McCabe, 2005; Eisenberg et al., 2019; Frey et al., 2017; Lönnqvist et al., 2015; Mata et al., 2018; Steiner & Frey, 2021).

Given the widespread adoption of self-report measures of risk preference in the behavioral sciences, it is surprising that there has been hardly any effort to systematically examine the cognitive processes and representations underlying people's self-reports (for exceptions, see Arslan et al., 2020; Jarecki & Wilke, 2018) and to thus shed some light onto the potential origins of these measures' desirable psychometric properties. Hence, several important questions remain largely unaddressed: What kind of evidence do people rely on during their judgment-formation process? What are the qualitative and quantitative properties of this process? And do cognitive explanations exist for the observation that people's self-reported risk preferences remain highly stable across time (e.g., Frey et al., 2017; Lönnqvist et al., 2015; Mata et al., 2018)? The goal of this article is to address these questions and "unpack" people's self-reports of their risk preferences, by modeling the information-integration processes underlying such self-reports, and thus testing the internal validity of this measurement approach.

The Psychology of Judgment Formation

Several streams in cognitive psychology assume judgment formation to rest on some form of *internal* or *external* information-sampling process. External information sampling (i.e., *Brunswikian sampling*; Fiedler & Juslin, 2005; Juslin & Olsson, 1997) has been extensively studied, often with the finding that observed samples predict people's choices and behaviors well (e.g., Fiedler, Renn, & Kareev, 2010; Hertwig, Barron, Weber, & Erev, 2006; Lindskog, Winman, & Juslin, 2013). In these investigations, computational models constitute a tool to systematically study the links between the external samples that participants observed and their choices or judgments—by formalizing the cognitive processes involved in information use and integration (e.g., Frey, Mata, & Hertwig, 2015; Frey, Rieskamp, & Hertwig, 2015; Kellen, Pachur, & Hertwig, 2016; Yechiam & Busemeyer, 2005).

When facing the task of providing a self-report, one typically cannot rely on external information but instead has to draw internal samples of one's own past behaviors and experiences (i.e., *Thurstonian sampling*; Bem, 1967; Fiedler & Juslin, 2005; E. J. Johnson, Häubl, & Keinan, 2007; Juslin & Olsson, 1997). This process may involve three broad stages, each involving different cognitive processes: First, information has to be retrieved from memory. Although memory retrieval has been a central assumption in models of *survey cognition* (e.g., Duckworth & Yeager, 2015; Jobe, 2000, 2003; Schwarz & Oyserman, 2001), concrete properties of this process have rarely been specified (for an overview, see Jobe & Herrmann, 1996; Koriat, Goldsmith, & Pansky, 2000; Tourangeau, Rips, & Rasinski, 2000). Second, the information retrieved from memory has to be integrated into an internal representation. Third and finally, the result of this information-integration process has to be rendered into a concrete output, for instance, mapping onto a specific response format (e.g., a Likert scale).

The focus of this article lies on the second stage; that is, the information-integration processes underlying people's self-reports. The respective cognitive processes have only rarely been studied systematically (for an exception, see Jarecki & Wilke, 2018) but information-integration processes are naturally paramount in research on judgment and decision making more generally (e.g., Dawes & Corrigan, 1974; Gigerenzer & Goldstein, 1996; Hastie & Park, 1986; Payne, Bettman, & Johnson, 1988). This line of research has identified three basic properties of evidence, which may also be of importance when people integrate information to render a self-report of their risk preferences: First, in their work on confidence judgments, Griffin and Tversky (1992) referred to the *weight of evidence* as the amount of information taken into account during a particular judgment (see also Kvam & Pleskac, 2016). Specifically, in their study on fairness assessments of biased coins (i.e., external samples), the weight of evidence referred to how many times a coin was spun and hence, the number of outcomes observed (i.e., sample size). Translated into the process of self-reporting one's risk preference, the weight of evidence may consist of how many pieces of information are retrieved from memory and either speak pro or contra risk taking. Indeed, initial evidence suggests that the weight of evidence of retrieved information may play an important role in the context of rendering self-reports (see introduction to study 1).

Second, the *strength of evidence* refers to the extremeness

of the available information; that is, how strongly a particular piece of information supports a certain judgment (Griffin & Tversky, 1992; Kvam & Pleskac, 2016; see also Koriat, 1993). In the work of Griffin and Tversky (1992), the strength of evidence was defined as how strongly the bias showed up across the entire sample of spun coins (i.e., effect size). Whereas in this example a single coin spin always yields equally strong evidence (i.e., in one or the other direction), in other contexts single pieces of information vary in terms of their strength of evidence (e.g., Hertwig & Pleskac, 2010). Translated into the process of self-reporting one's risk preference, this implies that a single yet "strong" piece of information may outweigh multiple "weak" pieces of information. The strength of evidence might play a focal role in the process of rendering self-reports, given the observations of Griffin and Tversky (1992) as well as Kvam and Pleskac (2016) that people tend to focus on the strength of evidence rather than on the weight of evidence when rendering judgments based on external samples. To date it remains untested to what extent the strength of evidence of available information is relevant in the context of rendering self-reports of risk preference.

Third, a large body of research into serial-position effects suggests that people are highly sensitive to the order of information (e.g., Hertwig, Barron, Weber, & Erev, 2004; Hogarth & Einhorn, 1992; Yechiam & Busemeyer, 2005). For example, it has been observed that the endowment effect may at least in part result from order effects in the aggregation process of respondents' internal samples, as information retrieved in the beginning was more indicative of participants' judgments (E. J. Johnson et al., 2007). Whereas some research into the sequential aggregation of internal or external samples has found such primacy effects (e.g., E. J. Johnson et al., 2007; Weber et al., 2007), other research suggests the occurrence of recency effects (e.g., Barron & Yechiam, 2009; Highhouse & Gallo, 1997; Hogarth & Einhorn, 1992). Hence, to the extent that people rely on multiple pieces of information when rendering self-reports, accounting for order effects may be important in the sense that information retrieved either at the beginning or at the end of the internal sampling process may be particularly influential.

Taken together, in contrast to research on decision making based on *external samples*, far less research exists on judgment formation based on *internal samples*—as are potentially drawn when rendering a self-report of one's risk preference. We aim to take a step towards closing this gap, by fostering a better understanding of the information-integration processes taking place when people render self-reports.

Aspect Listing: A Tool to Unpack Self-Reports

As information-integration processes typically remain hidden from direct observation, some research has employed the process-tracing method of *aspect listing* to gain a window into people's minds (E. J. Johnson et al., 2007; Weber et al., 2007; for a review see Schulte-Mecklenbeck et al., 2017). Specifically, this methodology entails prompting people to sequentially list their thoughts—typically referred to as *aspects*—that spontaneously cross their minds when responding to judgment or valuation questions. That is, by *not* prompting people to reflect on how they rendered a self-report *in hindsight*, this process-tracing method aims to avoid triggering any unnatural metacognitive processes, including potentially distorted post-hoc rationalizations (Nisbett & Wilson, 1977; but also see Hurlburt & Heavey, 2001).[1] Instead, and much like in research relying on think-aloud protocols, this method aims to trace the natural information-integration process "on the fly" (Ericsson & Simon, 1980).

Previous research adopting aspect listing has yielded several important insights, including into the cognitive processes underlying the endowment effect (e.g., E. J. Johnson et al., 2007), inter-temporal choice (e.g., Appelt, Hardisty, & Weber, 2011; Weber et al., 2007), the effect of attribute framing on choice (e.g., Hardisty, Johnson, & Weber, 2010), or domain-specificity in evolutionary content-domains (Jarecki & Wilke, 2018). Moreover, methods related to aspect listing ("thought-protocols" collecting information in a somewhat less structured way and after a judgment has already been provided) have also permitted several *qualitative insights* into judgment formation: For example, when rendering self-reports of life satisfaction (Schimmack, Diener, & Oishi, 2002) or risk preference (Arslan et al., 2020), people appear to rely mostly on personal experiences rather than on social comparisons (for details, see the introduction of study 1). Yet, contrary to the method of aspect listing these latter approaches do not readily permit the quantitative modeling of any information-integration processes, as the respective thought-protocols are typically not broken down into "atomic components" of evidence (e.g., the weight vs. strength of evidence).

Overview and Research Aims

The goal of this article is to promote a better understanding of the psychology underlying people's self-reports of their risk preferences: Although people are routinely asked to provide such self-reports in scientific and applied contexts, the underlying information-integration processes and people's respective cognitive representations remain largely unknown.

To this end, study 1 implemented a cognitive modeling approach to account for people's self-reported risk preferences

[1] Aspect listing may be complemented by prompting respondents to provide additional ratings of the aspects they had previously listed (e.g., how strongly an aspect speaks in favor of or against a particular choice or judgment), and such additional ratings would thus classify as a metacognitive task (Greifeneder & Schwarz, 2014; Hurlburt & Heavey, 2001; Koriat, 2007; Koriat et al., 2000).

based on various quantitative dimensions of evidence—as extracted from the listed aspects. Specifically, to what extent does a *cognitive account* potentially outperform various sociodemographic variables in predicting interindividual differences in self-reported risk preferences (RQ1a)? And how influential are the three reviewed properties of evidence in people's information-integration processes (RQ1b)? Moreover, by analyzing the content of the listed aspects, study 1 also permitted obtaining a range of qualitative insights into the cognitive representations of people's risk preferences (RQ2).

Subsequently, study 2 aimed at testing a longitudinal hypothesis that logically follows from the assumption that people's self-reports of their risk preferences emerge from quantifiable information-integration processes and robust cognitive representations. Specifically, the high temporal stability of self-reported risk preference, as observed repeatedly in previous research (e.g., Frey et al., 2017), may originate from relatively stable cognitive representations of one's own behaviors and experiences. Therefore, in a retest study we examined the stability of the content of the listed aspects (RQ3a) and the stability of the listed aspects' strength of evidence (RQ3b), to test whether stability and change in any of these two dimensions are systematically associated with stability and change in self-reported risk preference (RQ4a and RQ4b).

In addressing these research aims, we attached great importance to adhering to transparent and reproducible scientific practices and thus published a preregistration including the full theoretical rationale, all data, and the analyses scripts at https://osf.io/gndjw.

Study 1

Information-integration processes have long been of central interest in the literature on judgment and decision making (e.g., Dawes & Corrigan, 1974; Gigerenzer & Goldstein, 1996; Hastie & Park, 1986; Payne et al., 1988), and a diverse set of modeling approaches has thus emerged in this regard. For instance, in the framework of the Brunswikian lens model the cognitive integration of external cues into a judgment has been modeled descriptively by means of simple linear models (Hammond & Stewart, 2001; Hastie & Dawes, 2001), which were also used to address normative questions concerning information integration in various judgment processes (e.g., the role of proper vs. improper linear models in decision making; Dawes, 1979). Another substantive body of research has focused on noncompensatory heuristics to examine information use and integration in the context of inferential choice (e.g., take-the-best, TTB; Gigerenzer & Brighton, 2009; Gigerenzer & Goldstein, 1996, 1999). Finally, research into sequential information integration has developed sophisticated fractional-adjustment models to study, for instance, the role of serial position effects in information integration (e.g., Hogarth & Einhorn, 1992; Sutton & Barto, 1998).

To address RQ1 in study 1—that is, how well people's self-reported risk preferences can be quantitatively accounted for based on the listed aspects (RQ1a), and how influential different properties of these aspects are in people's self-reports (RQ1b)—we built on these different strands of research and implemented a twofold-approach. First, we directly sampled a set of models from the literature, aimed at covering a large model space to thus incorporate models that account for (different combinations of) the strength of evidence, the weight of evidence, and the order of evidence. As reviewed above, these dimensions constitute three key properties of evidence that people may rely on when rendering self-reports. Our approach followed a proof-of-concept provided by Jarecki and Wilke (2018), who used cognitive process models to study risk taking in different evolutionary domains. Yet, our approach was different in the sense that it focused on general risk preference, modeled continuous self-reports (as opposed to hypothetical binary choices), and importantly, took into account the listed aspects' strength of evidence—a property that may be highly relevant during information integration according to previous observations (Griffin & Tversky, 1992; Kvam & Pleskac, 2016). Second, as some of the different models turned out to yield similar predictions when applied to the empirical data of study 1 we also implemented a set of Bayesian ordinal regression models using the listed aspects' weight of evidence and strength of evidence as direct predictors of self-reported risk preference—to thus facilitate a direct comparison between the roles of these two properties of evidence.

Beyond analyzing quantitative aspects of the information-integration processes (i.e., RQ1a and RQ1b), the method of aspect listing also permitted conducting a series of more qualitative analyses, which allowed insight into people's cognitive representations of their risk preferences (RQ2). Specifically, these analyses characterized the content and sources of the aspects people rely on during information integration, such as whether people predominantly tap into personal experiences or social comparisons to render their self-reports (e.g., Arslan et al., 2020; Schimmack et al., 2002), or how frequently people typically experience in daily life what they consider as aspects during judgment formation. Previous research along these lines, which has prompted respondents to *explain* their previously stated risk preferences, found that people mainly considered risks that they had personally taken, which were rather voluntary, had known and controllable consequences, and were old and familiar (Arslan et al., 2020). Our approach promised to corroborate and extend these findings, as we prompted participants to *concurrently* list the aspects that crossed their minds *during judgment formation* (i.e., as opposed to after already having provided a response, which in principle could lead to distorted reports; Nisbett & Wilson, 1977). Moreover, as the approach

of aspect listing taps into people's cognitive representations in a semi-structured way (i.e., collecting aspects one by one), it is possible to examine, for instance, the content and sources separately for aspects that speak either pro or contra risk taking (i.e., pro-aspects and contra-aspects, respectively). In addressing RQ2, we again relied on a two-fold approach: We first analyzed the cognitive representations based on respondents' own ratings of their aspects. Second, we also relied on the evaluations of a subset of 300 aspects as provided by external raters. The latter evaluations rendered possible further insight concerning the content of the aspects (e.g., classification to various content domains) as well as an external validation of the aspects' strength of evidence.

Methods

We collected data from 250 participants via Amazon MTurk (115 females; mean age: 37.4 years; range: 18 – 73 years; mean number of years of education: 15.2; modal income: 1,000 - 2,000 USD per month). To ensure a high data quality, only MTurkers with an approval rate of at least 95% and who had completed at least 500 HITs (i.e., human intelligence tasks) on Amazon MTurk were eligible to participate (Peer, Vosgerau, & Acquisti, 2014; see also Buhrmester, Kwang, & Gosling, 2011; Casler, Bickel, & Hackett, 2013; Paolacci, Chandler, & Ipeirotis, 2010). Moreover, participants had to pass two attention check questions and provide ratings of at least 25 out of 100 on questions asking how focused they were and how much effort they put into the study. Data were collected in 2019. Study completion on average took 6 minutes, for which participants were reimbursed with 0.85 USD. Both studies were approved by the ethics committee of the Faculty of Psychology of the University of Basel (#023-18-1).

According to a prior model recovery analysis (see preregistration; cf. Gluth & Jarecki, 2019), a sample of 250 participants was sufficiently large for the separate models to be recovered with high recovery rates, except for two models which were thus excluded from the model space. This sample size is also sufficient to detect small to medium effects in a frequentist framework (f^2 of .03 with a power of $1 - \beta = .80$; calculated using G^*Power 3.1, Faul, Erdfelder, Buchner, & Lang, 2009) for the most complex regression model involving three predictors (i.e., query theory, see below). Note, however, that we conducted all analyses in a Bayesian framework and we thus report 95% credible intervals (95% CIs) rather than p-values (unlike in a frequentist framework, the 95% CI indicates the range that contains the population parameter with a probability of 95%).

All analyses were performed using *R* (R Core Team, 2020). We used the *rstanarm* and *brms* packages for the regression analyses (Bürkner, 2017; Goodrich, Gabry, Ali, & Brilleman, 2018) and implemented the default priors as provided by these packages (see Supplemental Material, SM, section 4.3).

Procedure

After reading general instructions, participants provided informed consent and sociodemographic information. They were then shown the general risk item of the SOEP ("Are you generally a person who is willing to take risks or do you try to avoid taking risks?"; e.g., Dohmen et al., 2011). Yet, prior to actually providing an answer participants were prompted to think of (and list) all reasons that crossed their minds *while* coming up with an answer (the exact wording is available on https://osf.io/gndjw). Specifically, participants had to report at least one aspect and were asked to continue reporting aspects until they could not think of any further aspects. Once done with this task, participants provided their rating to the SOEP general risk item on a scale ranging from 0 to 10. The procedure of first implementing the aspect listing followed the original protocol (E. J. Johnson et al., 2007; Weber et al., 2007); we specifically pretested potential order effects (i.e., self-reported risk preference first vs. aspect listing first) in a dedicated pilot study (see preregistration), which revealed no credible mean differences of self-reported risk preferences in the two examined orders. Finally, participants were sequentially presented with the aspects that they had previously listed (in randomized order), and were prompted to evaluate these aspects on a series of dimensions (i.e., including the aspects' strength of evidence as well as dimensions tapping the content and sources of the aspects; see Table 1 and the respective sections below).

Procedure and analyses concerning RQ1a and RQ1b

To formally model the information-integration processes underlying people's self-reports (RQ1a and RQ1b), we implemented the following steps.

Operationalization of the aspects' properties of evidence. We operationalized the three quantitative properties of evidence reviewed above as follows (see Table 1): First, we operationalized the strength of evidence as participants' ratings of how strongly each aspect supports risk-avoidance or risk-seeking, ranging from -50 to 50. Second, to determine the weight of evidence, we classified the aspects—based on the rated strength-of evidence—as either pro-aspects (strength of evidence > 0) or contra-aspects (strength of evidence < 0) and then counted the number of pro- and contra-aspects for each participant. Third and finally, the order of evidence naturally followed from the sequence by which participants listed the aspects.

Model space and model selection criteria. We initially implemented six separate models (for a detailed description of all models, see SM section 4.1) to cover various combinations of the three properties of evidence as reviewed above. Specifically, the EXT model (inspired by the TTB heuristic; Gigerenzer & Goldstein, 1996, 1999) used the most extreme

strength of evidence (i.e., the one the furthest away from the center of the scale) as predictor; the FIRST model used the strength of evidence of the aspect listed first in the sequence as predictor (for related lexicographic models such as take-the-first; see Jarecki & Wilke, 2018; J. G. Johnson & Raab, 2003); and the LAST model used the strength of evidence of the aspect listed last in the sequence as predictor. These three models were non-compensatory models; the remaining three models were compensatory models. Specifically, the SUM model (a weighted additive model; see Payne et al., 1988) used the sum of the strength of evidence of all aspects listed by a participant as predictor; query theory (QT; E. J. Johnson et al., 2007; Weber et al., 2007) was implemented as a linear model with the weight of evidence (the number of pro-aspects and the number of contra-aspects separately) and the order of evidence as predictors; and finally, the value updating model (VUM; an instance of a fractional-adjustment model; Hertwig et al., 2006; Hogarth & Einhorn, 1992) implemented a weighted average of the strength of evidence as predictor, rendering possible the capture both of primacy and recency effects.

To enable a fair model comparison—accounting for the fact that some models (i.e., QT and VUM) had free parameters whereas others did not—we purely focused on predictive accuracy (i.e., out-of-sample prediction). Thereby, adjustable parameters only provide an advantage for a model if they actually help explain systematic variance (e.g., Yarkoni & Westfall, 2017). To this end, we employed a five-fold cross-validation approach. That is, we partitioned the data in five subsets (folds) and used four folds to fit the free parameters (i.e., in the case QT and VUM) and predicted the fifth (hold-out) fold with the obtained parameter estimate. Given our data structure with one response per participant, all parameters were estimated across participants. This procedure was repeated for all models until each of the five folds was predicted once by every model. We then determined the average (i.e., across the independent hold-out samples) Spearman rank correlations (i.e., r_s) between the model predictions and the self-reported risk preferences.

Based on a prior model recovery analysis (see preregistration), the six initial models were expected to yield somewhat correlated yet sufficiently distinguishable predictions. Ultimately, however, in study 1 the models ended up making relatively similar predictions, given the average of 3.4 aspects that participants listed (note that this number is in line with previous studies, e.g., Jarecki & Wilke, 2018; E. J. Johnson et al., 2007; Weber et al., 2007), and given that participants tended to list either only pro-aspects or only contra-aspects (which also made it difficult to systematically study order effects). To illustrate, FIRST and LAST resulted in very similar model predictions, and a parameter recovery analysis for the VUM indicated that different values for the weighting parameter (i.e., capturing recency or primacy effects) resulted in very similar model predictions (see SM section 4.5).

We thus also pursued a complementary approach as a robustness check. Specifically, we employed two Bayesian ordinal regression models, using the aspects' strength of evidence and weight of evidence (i.e., averaged per participant; when averaging the weight of evidence each pro-aspect was given the value 1, and each contra-aspect was given the value -1), respectively, to predict self-reported risk preferences (see Table 1). We relied on multiple indices to compare these models: First, we compared their expected log predictive density (ELPD)—a statistic that provides an estimate of the to-be-expected out-of-sample predictive performance—based on the leave-one-out information criterion (LOOIC), which is similar to the Akaike information criterion (AIC) but better suited for Bayesian model comparisons, as it can account for the implemented priors (Vehtari, Gelman, & Gabry, 2017). Moreover, we compared the models' accuracies, their chance corrected accuracies (Cohen's κ), as well as how often their (correct) predictions coincided in a tournament approach (see Broomell, Budescu, & Por, 2011).

Reference models. To compare the described models against a baseline, we also implemented three Bayesian ordinal regression models that inferred participants' self-reported risk preferences based on up to five sociodemographic predictors. The first model included age as the sole predictor, the second model included sex as the sole predictor, and the third model included age, sex, years of education, income, and employment status as predictors. These variables have been suggested to be systematically associated with individual differences in risk preference, and in the case of age (e.g., Mamerow, Frey, & Mata, 2016; Mata, Josef, & Hertwig, 2016) and sex (e.g., Byrnes, Miller, & Schafer, 1999), these associations were found to be particularly robust (for an overview, see Frey, Richter, Schupp, Hertwig, & Mata, 2020). To compare these reference models with the other two ordinal regression models, we included them in the tournament approach described above, and additionally relied on the LOOIC-based ELPD. Finally, to compare the reference models with the initial set of models described above, we also report Spearman correlations between the model predictions and participants' self-reported risk preferences.

Procedure and analyses concerning RQ2

To examine the content and sources of the aspects people rely on during judgment formation (RQ2), we implemented the following steps.

Participants' own ratings of their aspects. At the end of the study, participants provided ratings of each aspect they had previously listed, concerning (a) how strongly the aspect supported risk seeking versus risk avoidance (i.e., the strength of evidence used in the modeling analysis; see above), (b) whether the aspect included a previous personal experience (see Arslan et al., 2020), (c) whether the aspect

included a comparison with another person (see Arslan et al., 2020; Schimmack et al., 2002), (d) how often participants typically experience in their daily lives what they described in the aspect (i.e., relatively common or rather rare but potentially high-stake events; see Hertwig et al., 2004), (e) whether the aspect referred to an active choice or a passive experience (i.e., voluntary or involuntary exposure to risks; Fischhoff, Slovic, Lichtenstein, Read, & Combs, 1978), and (f) whether the aspect involved something controllable or uncontrollable (see Arslan et al., 2020; Fischhoff et al., 1978; Maccrimmon & Wehrung, 1985). Table 1 provides an overview and a detailed description of the items used.

For each of these dimensions, we provide the distributions of participants' ratings, separately for pro- and contra-aspects, and report post-hoc mixed-effects models to explore any systematic differences between pro- and contra-aspects. Specifically, we ran generalized linear mixed-effects models predicting the various ratings and using the aspects' direction (pro or contra) as dummy coded predictors, using by-subjects random slopes and intercepts (Barr, Levy, Scheepers, & Tily, 2013). We also quantified the differences in the sentiment for pro- and contra-aspects (see SM sections 2 and 4.2).

External ratings of a subset of 300 aspects. For 300 randomly selected aspects (i.e., about one third of the 857 aspects listed in study 1), we also collected external ratings from three independent raters (i.e., the first author and two research assistants; using a majority rule to integrate the three ratings; see SM section 5.5 for further methodological details).

First, the raters inferred the listed aspects' strength of evidence, to thus provide an independent validation of participants' own ratings. To this end, we provided the same scale as participants used to evaluate their own aspects.

Second, the raters assessed a range of additional properties that were not assessed by participants themselves. These properties stem from five risk categories and have been suggested to be important drivers of and motives underlying risk-taking behaviors, covering both stable dispositions (i.e., traits) as well as situational characteristics (i.e., state variables), namely: (a) outcome-related properties (e.g., the magnitude of the positive outcomes; Kahneman & Tversky, 1979; Sitkin & Pablo, 1992), (b) goal/state-related properties (e.g., whether the goal was to keep or improve one's status quo; e.g., Lopes, 1984; Mishra, Barclay, & Sparks, 2017), (c) properties related to cultural roles and personality (e.g., whether a social norm or one's personality was mentioned; Nicholson, Soane, Fenton-O'Creevy, & Willman, 2005; Sitkin & Pablo, 1992), affect-related properties (e.g., whether a feeling of fear or thrill was mentioned; Lerner, Gonzalez, Small, & Fischhoff, 2003; Loewenstein, Weber, Hsee, & Welch, 2001; Zuckerman, 2002), and (d) properties related to life-history (e.g., whether one's age or children were mentioned; e.g., Wang, Kruger, & Wilke, 2009). Finally, we used an *other* category to classify whether an aspect just relativized (e.g., "that depends on the situation"), or only contained semantically invalid sequences of letters. Please see SM section 5.5 for the complete list of properties along with some key references and the full description of the rating procedure.

Third, the raters inferred the life domains to which the listed aspects supposedly belong to. To this end we provided the domains as suggested in one of the most popular domain-specific risk-taking questionnaires (DOSPERT; Blais & Weber, 2006; Frey, Duncan, & Weber, 2020; Weber, Blais, & Betz, 2002), those suggested in the SOEP (e.g., Dohmen et al., 2011), as well as those of the evolutionary risk scale (ERS; Wilke et al., 2014)—overall resulting in 19 different domains (see SM section 5.5).

Results

In line with previous observations, the self-reports of the majority of participants (57%) indicated risk-aversion (i.e., most participants provided a rating of lower than five on the scale ranging from 0 to 10), with an average rating of M = 4.2. The majority of participants (81%) listed between one and four aspects (M = 3.4; range: 1 – 12). Matching participants' overall tendency for risk-aversion, the majority of these aspects were contra-aspects (61%). Moreover, most participants (82%) only listed either contra-aspects or pro-aspects, directionally matching their risk preference (i.e., risk-seeking vs. risk-averse). Participants' ratings of their aspects' strength of evidence were relatively consistent within participants, with an intra-class correlation of .76.[2]

Our validation of the aspects' strength of evidence using external raters showed a high degree of agreement: The strength of evidence as assessed by the external raters (i.e., average across the three raters) and the strengths of evidence as indicated by participants themselves correlated with r_s = .82. Moreover, in 93% of the cases the three raters classified the listed aspects correctly (i.e., in line with participants' own judgments) as pro- or contra-aspects.

RQ1: Modeling self-reported risk preferences

As outlined above, we followed a two-fold approach to modeling self-reported risk preferences. First, we compared six separate models directly sampled from the literature on judgment and decision making. These models were capable of predicting self-reported risk preference well, with r_s ranging from .78 to .90; note that these values resulted from out-of-sample predictions using the independent hold-out sets. Specifically, the correlations between model predictions and actual self-reports were r_s = .90 (VUM), r_s = .84 (LAST),

[2]This analysis was run with an intercept-only model, with by-subjects random intercepts predicting the aspects' strength of evidence.

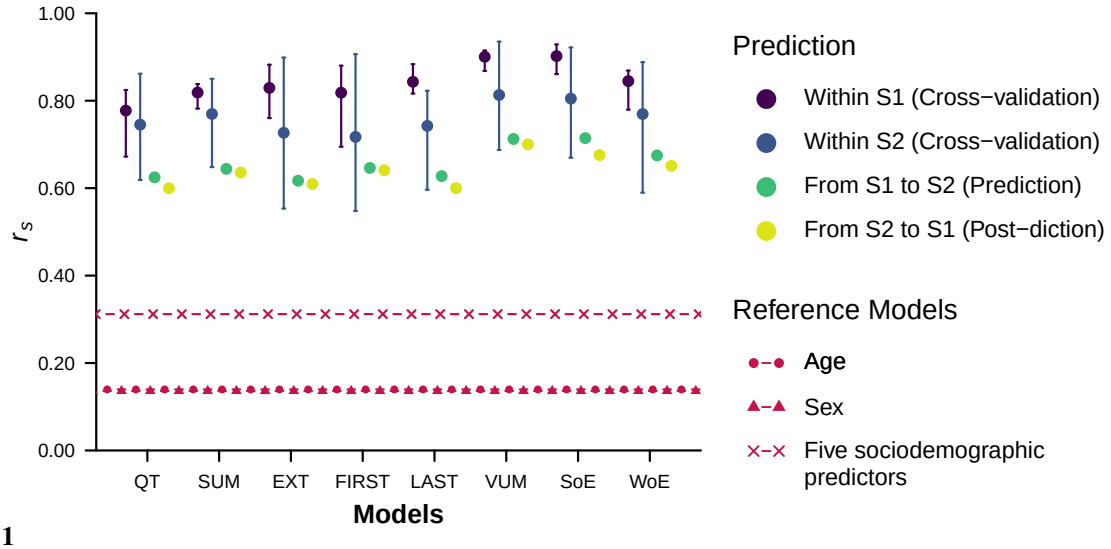

Figure 1

Spearman correlations between the different model predictions and self-reported risk preference. QT = Query theory; SUM = Sum of evidence; EXT = Most extreme evidence; FIRST = First aspect's evidence; LAST = Last aspect's evidence; VUM = Value updating model. SoE = Ordinal regression model with the average strength of evidence per participant as predictor. WoE = Ordinal regression model with the average weight of evidence per participant as predictor. Whiskers depict the range of r_s in the five folds of the cross-validation within studies. For the pre-/post-diction across studies, the models only used the aspects participants listed in one study to pre-/post-dict their risk preferences in the other study. "Five sociodemographic predictors" = Reference model using age, sex, years of education, income, and employment status as predictors. All reference models were implemented separately for study 1 and study 2 and their respective r_ss averaged for this plot.

r_s = .83 (EXT), r_s = .82 (FIRST), r_s = .82 (SUM), and r_s = .78 (QT). Moreover, these models clearly outperformed the three reference models, with correlations between the predictions of the latter and the self-reports ranging from r_s = .14 to r_s = .31 (see Figure 1).

Second, we compared a set of ordinal regression models using the strength of evidence (SoE) and the weight of evidence (WoE) as direct predictors. Corroborating the results reported above, both models performed well, with r_s = .90 (SoE) and r_s = .85 (WoE). Moreover, the model including the strength of evidence as predictor outperformed the model including the weight of evidence as predictor by eight percentage points of *correct predictions* (see Table 2). Also, there was robust evidence that the strength of evidence was a more important predictor than the weight of evidence according to the direct model comparison based on the two models' to-be-expected out-of-sample predictive performance (i.e., LOOIC-based ELPDs; see Table 2).

Following the tournament approach proposed by Broomell et al. (2011), we also gauged the proportion of identical model predictions of the five ordinal regression models (i.e., the two models using the strength and weight of evidence as predictors, and the three reference models). While some models resulted in highly similar predictions (i.e., the reference models including only age or sex as predictors made identical predictions in 97% of the cases), the two models using the different properties of evidence as predictors were sufficiently distinguishable (see Table S2 and Figure S5; see also Table 2).

Finally, again in line with the comparison of the models reported above, both tested properties of evidence proved to be better predictors than any of the reference models that used sociodemographic predictors. Specifically, the strength of evidence model outperformed the best reference model by 18 percentage points, and the weight of evidence model outperformed the best reference model by ten percentage points (see Table 2).

RQ2: Sources and content of the listed aspects

Our analyses of people's cognitive representations of their risk preferences (see Figure 2) indicated that most participants retrieved personal experiences (and less so social comparisons) when rendering their self-reports (more so for pro- than contra-aspects: b = 1.87, 95% CI: [0.82, 3.17]). Furthermore, the listed aspects involved mostly active choices rather than passive experiences (more so for pro- than contra-aspects: b = 1.44, 95% CI: [0.81, 2.22]), and situations with rather controllable outcomes (no credible differences between pro- and contra-aspects: b = 0.24, 95% CI: [-0.52, 1.08]). Across the listed aspects, participants' answers to

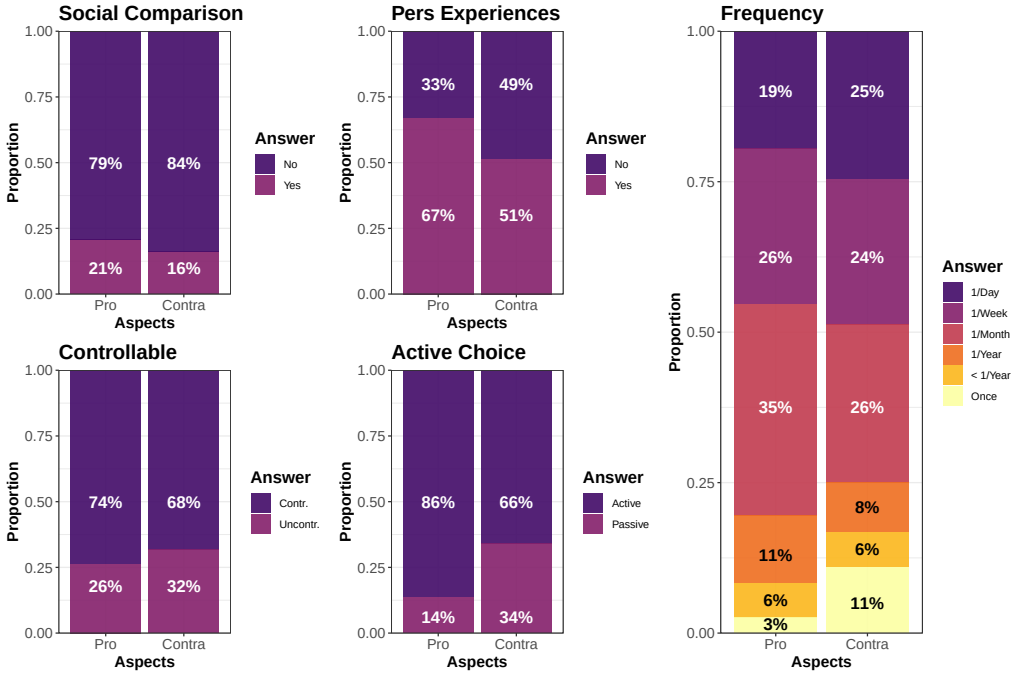

Figure 2

Distributions of the sources and content of the listed aspects from study 1 across all participants and aspects.

these questions were quite consistent; that is, most participants rated their respective aspects similarly on a given question. Furthermore, the listed aspects were typically not rare situations or experiences, but frequent encounters in participants' daily lives (i.e., the categories once per day, once per week, and once per month made up for 80.4% and 74.8% of all pro- and contra-aspects, respectively). Finally, most aspects had a negative sentiment (see SM section 2), but the pro-aspects less so than contra-aspects ($M_{pro-aspects}$ = -0.52; $M_{contra-aspects}$ = -1.12; b = 0.60, 95% CI: [0.40, 0.78]).

As Figure 3 illustrates, positive emotions and feelings as reflected by the words *fun* or *enjoy* often occurred in pro-aspects, along with words describing positive outcomes such as *reward*, *gain*, or *benefit*. The picture looked substantially different for contra-aspects, where *lose*, *money*, or *safe* where very prominent mentions, along with negative emotions or feelings expressed by words such as *hurt*, *afraid*, or *worry*.

Our additional analyses using external ratings of the listed aspects showed that participants mostly retrieved domain-general statements (79.8%), and if domain-specific statements were retrieved, these were mostly in the domains of health/safety (9.8%), financial (7.1%), social (4.0%), occupation (3.4%), recreational (2.7%), and kinship (2.0%).[3] In the SM (section 5.5) we report an analysis showing that the domains put forth by two established domain-specific risk-taking scales could be recovered well——that is, in 13 out of the 15 distinct domains suggested by these two scales, more than half of the respective items were correctly recovered, and in eight of the 15 domains all items were correctly recovered. Regarding the potential drivers and motives underlying risk taking, we found that participants mostly considered the valence of the potential outcomes (i.e., positive outcomes in the pro-aspects, but also often in combination with negative outcomes—i.e., indication of a risk-return trade-off; and more negative outcomes in the contra-aspects). Moreover, participants often mentioned their positive (in the case of pro-aspects) and negative feelings (in the case of contra-aspects) towards taking risks. Finally, in pro-aspects participants often adopted an opportunity focus, aimed at improving their status quo, while in contra-aspects, participants often adopted a safety focus, aimed at keeping their status quo (see also Figure S9).

Discussion

The listed aspects—and more precisely, different properties of evidence thereof, particularly the strength of evidence—turned out to be highly predictive of participants' self-reported risk preferences: Overall the cognitive modeling approach performed substantially better than using a series of sociodemographic indicators as predictors (i.e., reference models), which suggests that people recruit system-

[3]The external raters could select multiple domains per aspect, which is why these numbers do not add up to 100%.

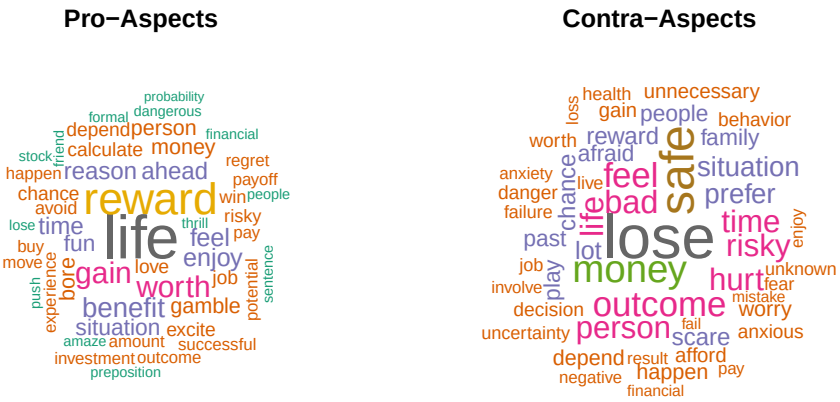

Figure 3

Word cloud of the most frequently occurring words, shown separately for pro- and contra-aspects. Larger fonts indicate higher frequencies of occurrence of the respective words in the aspects listed in study 1. The words risk*,* take*, and* avoid *were excluded to increase the visibility of the other words that occurred less frequently than these words that were also part of the formulated question and thus were often repeated in the aspects (e.g., "I avoid taking risks because..."). The words listed in the word cloud of pro-aspects had a clearly positive average sentiment (M = 7.07), and those listed in the word cloud of contra-aspects had a clearly negative average sentiment (M = -3.60)*

atic information-integration processes when rendering self-reports of their risk preferences. Crucially, external ratings of the aspects' strength of evidence were closely aligned with participants' own ratings, thus making it unlikely that the high predictive power of this property of evidence simply arose due to the close temporal proximity between the respective ratings (note that we also conducted an independent cross-study analysis as a further robustness check in this regard, see below).

Although we faced some constraints when comparing the six initial models sampled from the literature—particularly concerning the role of the order of evidence—a direct evaluation of the relative importance of the aspects' strength of evidence and weight of evidence yielded clear and corroborating results. Furthermore, given the structure of our data it may seem somewhat surprising that the non-compensatory models were outperformed by the compensatory models (see Gigerenzer & Gaissmaier, 2011; Gigerenzer & Goldstein, 1996), yet it is important to keep in mind that these differences in model performance were small.

Participants' cognitive representations of their risk preferences proved to rest mostly on situations that involved active choices rather than passive experiences, suggesting that most people may think of risk taking as an explicit decision. The mostly domain-general information retrieved by participants tended to focus on the valence of the outcomes, often referring to explicit trade-offs in line with a risk-return framework (Weber et al., 2002; Weber & Milliman, 1997). In pro-aspects, participants often expressed an opportunity focus, whereas in contra-aspects they often expressed a safety focus. Moreover, in line with some conceptualizations of risks (e.g., Bell, 1982; Loewenstein et al., 2001; Loomes & Sugden, 1982; Mellers, Schwartz, Ho, & Ritov, 1997), many aspects mentioned positive or negative feelings.

Taken together, the results of study 1 suggest that when people are prompted to report their own risk preferences, they may retrieve information from memory and evaluate *how strongly* multiple pieces of information support a specific judgment (i.e., strength of evidence). Thus, on a quantitative level the information-integration processes in the context of evaluating one's own risk preference appear to share similarities with those of evaluating external objects (Griffin & Tversky, 1992; Kvam & Pleskac, 2016).

Study 2

In study 2 we tested a longitudinal hypothesis that logically follows from the basic assumption that people's self-reported risk preferences are robustly rooted in their cognitive representations of idiosyncratic experiences and behaviors. Specifically, in RQ3 we tested whether two particular dimensions of the retrieved information show stability across time—that is, from study 1 to study 2—namely, (a) whether people retrieve the identical aspects (aspect stability) and (b) whether the listed aspects have, on average, a similar strength

of evidence (evidence stability). To illustrate, to the extent that people sample aspects from a large pool of idiosyncratic experiences, they may not necessarily retrieve the exact same aspects at different occasions (e.g., because different contexts may prime the retrieval of a particular type or class of aspects)—yet this naturally does not preclude the possibility that the retrieved aspects still suggest a similar degree of risk preference. Consequently, RQ4 examined whether the stability of self-reported risk preferences directly hinges on aspect stability or on evidence stability.

Methods

Of the 250 participants in study 1, 164 accepted an invitation to complete a retest study after an interval of one month. Of these participants, 150 passed all quality checks and their data were used for the subsequent analyses (72 females; mean age: 39.07; range: 19 – 70 years; mean number of years of education completed: 15.44; modal income: 2,000 - 3,000 USD per month). We deviated from our preregistered analysis plan on four minor points (see SM section 3).

The participants who completed both studies did not differ credibly from participants who only completed study 1 in terms of their self-reported risk preferences, average strength of evidence of the listed aspects, average sentiment of the listed aspects, years of education, or the proportion of females (see SM section 5.6); however, the former participants tended to be slightly older (b = 2.90, 95% CI [0.16, 5.40]) and on average listed slightly more aspects (b = 0.73, 95% CI [0.26, 1.15]). In sum, if at all there were only very weak indications for systematic selection effects.

Procedure

The design of study 2 was equivalent to that of study 1, with the exception that we added two questions at the end of the study. Specifically, we asked participants how well they could remember the aspects they had listed in study 1, as well as concerning their intuition of how similar their listed aspects were across studies. Both of these ratings were provided on a scale ranging from 0 to 100. Participants again received compensation of 0.85 USD for their participation.

External similarity ratings to gauge aspect stability

To examine aspect stability, we first obtained similarity ratings for the listed aspects. To this end, we asked 63 independent raters (recruited via Amazon MTurk) to judge the similarities of all possible pairs of aspects that were listed by each participant across and within the two studies. Pairs of aspects were partitioned into packages of about 200, and for each package three raters were asked to provide their judgments using a Likert scale ranging from 0 to 5 (i.e., each rater rated a total of around 200 aspect pairs, one pair at a time). To gauge the inter-rater agreement we calculated Kendall's coefficient of concordance (W; Kendall, 1948) for each triplet of raters who evaluated the similarities of the same aspects (M_W = .56, range = .38 - .77).

We denoted two aspects to be "equivalent" using a very conservative cutoff of five (i.e., mean similarity rating across the three raters, implying that all raters had to provide the highest rating). To obtain the proportion of equivalent aspects we divided the number of equivalent aspects by the maximal number of aspects that could be equivalent; across studies, the maximally possible number of equivalent aspects is equal to the smaller number of aspects listed in study 1 and study 2. As a robustness check, we also used additional ways to aggregate similarity ratings in the analyses concerning aspect stability (SM section 5.4).

Statistical analysis

To quantify the relation between aspect stability and the stability of the self-reported risk preferences (RQ4a), we used a gamma regression model with a log link function. This allowed us to account for the skewness in the absolute difference scores of the self-reported risk preferences. For the robustness test with the average similarity rating as predictor, we again used a gamma regression model with a log link function.

To quantify the relation between evidence stability and the stability of the self-reported risk preferences (RQ4b), we used a linear regression model with both the evidence stability and the self-reported risk preferences scaled for better interpretability. In contrast to the relation between overlaps and change in the self-reported risk preferences in RQ4a, the variables involved in RQ4b—that is, the change in the aggregated strength of evidence and the change in the self-reported risk preferences—allow for testing a directional relationship. Therefore, we did not use the absolute differences but the directional difference scores of the variables between study 1 and study 2. We again used the default priors implemented in *rstanarm*.

Results

Just as in study 1, the self-reports of the majority of participants (60%) indicated risk-aversion, with an average rating of M = 3.81. The majority of participants (78%) again listed between one and four aspects (M = 3.6; range: 1 – 13), and within participants the number of listed aspects was quite similar from study 1 to study 2 (r_s = .54). On average, participants indicated that they did not actively remember the aspects they had listed in study 1 (M = 21.56, SD = 24.89; on a scale from 0 to 100). Nevertheless, participants appeared to have an intuition that the aspects they had listed in study 1 were rather similar to those they had just listed in study 2 (M = 65.42, SD = 22.05; again on a scale from 0 to 100). The strength of evidence of the listed aspects again proved to

be the most important predictor of participants' self-reported risk preferences (see Figure 1).

Finally, in line with previous observations (e.g., Frey et al., 2017; Mata et al., 2018) participants' self-reported risk preferences were highly stable at a one-month interval (r_s = .80). The first panel of Figure 4 depicts the distribution of within-subject differences, which is clearly centered on zero.

RQ3: Aspect stability and evidence stability

We examined aspect stability by determining the proportion of equivalent aspects across studies (see methods section). With the strict criterion imposed for classifying aspects as equivalent, aspect stability was relatively low: Only every twentieth aspect pair (i.e,. a proportion of .05) fulfilled the criterion of equivalence.

Yet, the picture was substantially different for evidence stability: Specifically, the strength of evidence (aggregated over all aspects listed by each participant)[4] remained highly stable across time (r_s = .68), as can be seen in the second panel of Figure 4. A Bayesian paired t-test corroborated that there was no credible difference between the aggregated strength of evidence of a participant's aspects listed in the two studies (Δ_M = -1.39, 95% CI [-3.48, 0.64]).

RQ4: Relationship of the stability of self-reported risk preference with aspect stability and evidence stability

Aspect stability was not credibly associated with the stability of self-reported risk preference (b = -1.03, 95% CI [-2.90, 1.59]; r_s = -.12) nor did a robustness test (i.e., using the average similarity ratings instead of the proportion of equivalent aspects) indicate a credible association between aspect stability and the stability of self-reported risk preference (b = -0.43, 95% CI [-0.95, 0.10]; r_s = -.20).

Conversely, and as can be seen in the third panel of Figure 4, evidence stability across the two studies was credibly and strongly associated with the stability of self-reported risk preference (β = 0.63, 95% CI [0.50, 0.75]; r_s = .45).

Discussion

Study 2 corroborated the results obtained in study 1, and replicated previous observations of a high temporal stability of self-reported risk preference (e.g., Frey et al., 2017; Mata et al., 2018). More importantly, our analyses revealed that within participants the listed aspects' average strength of evidence remained highly stable across the two studies; that is, although participants did not necessarily list the exact same aspects across the two studies (low aspect stability), they appeared to have sampled and listed aspects from a pool of idiosyncratic experiences with comparable strength of evidence (high evidence stability). Crucially, changes in the strength of evidence were systematically associated with changes in self-reported risk preferences. In sum, our analyses suggest that people's internal sampling process results in the retrieval of aspects that yield high evidence stability—thus providing a cognitive explanation for why self-reported risk preferences remain stable across time.

Cross-Study Analysis

Finally, to further clarify the predictive power of the strength and weight of evidence of the listed aspects, we repeated the analyses reported in study 1 by focusing on *cross-study pre- and post-dictions*. These analyses were particularly targeted at ruling out the possibility that the high predictive power of the aspects' strength and weight of evidence resulted from a methodological artifact; namely, that the respective ratings were provided in close proximity to the self-reported risk preferences. Thus, being able to predict (i.e., from study 1 to study 2) and post-dict (i.e., from study 2 to study 1) participants' self-reports of their risk preferences only using the aspects listed in the other study would constitute substantial evidence for the robustness of our main findings. Naturally, these tests rest on the assumption that people's risk preferences remain at least somewhat stable across time, a finding that has repeatedly been documented (e.g., Frey et al., 2017; Mata et al., 2018).

Methods

Just as in study 2, we relied on the data of the 150 participants who completed both studies for this cross-study analysis. We again implemented the two-fold approach used in study 1; that is, we performed the cross-study analyses both with our initial set of six models, as well as with the ordinal regression models. To this end, we relied on the aspects (and estimated model parameters) obtained in study 1 (study 2) to generate predictions for the self-reported risk preference of study 2 (study 1).

Results

As can be seen in Figure 1, in the cross-study analyses the predictive accuracies of the six initial models were still substantial, with r_s ranging from .60 to .71. Specifically, the correlations between model predictions (from study 1) and self-reports (in study 2) were r_s = .71 (VUM), r_s = .65 (FIRST), r_s = .64 (SUM), r_s = .63 (LAST), r_s = .63 (QT), and r_s = .62 (EXT). Moreover, the correlations between model post-dictions (from study 2) and self-reports (in study 1) were r_s = .70 (VUM), r_s = .64 (FIRST), r_s = .64 (SUM), r_s = .61 (EXT), r_s = .60 (LAST), and r_s = .60 (QT).

[4] In line with our preregistered analysis plan, we used the value updating model to aggregate evidence stability across the aspects listed by each participant, because it was the best-performing of the original models in both studies. Yet, using the arithmetic mean to aggregate the strength of evidence yielded an equivalent result (r_s = .67).

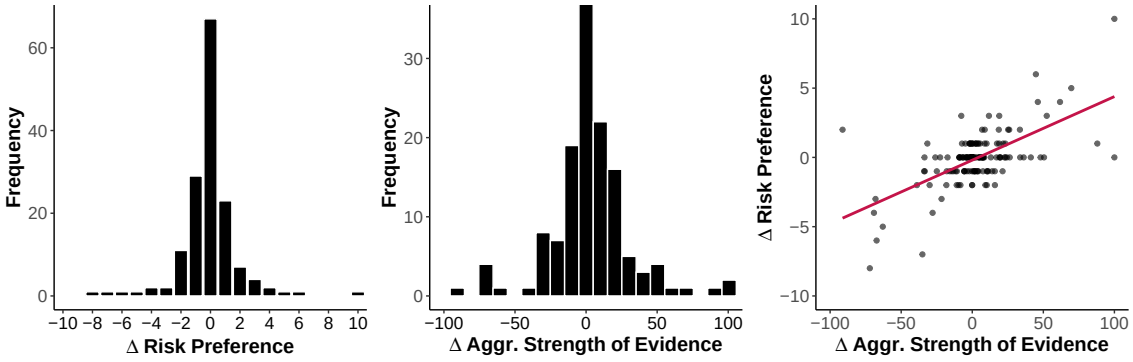

Figure 4

Stability of self-reported risk preference (first panel) and stability of the aspects' strength of evidence (second panel). The histograms show the distributions of within-subject differences between study 1 and study 2. The relation between changes in the aggregated strength of evidence and changes in self-reported risk preference is shown in the third panel (r_s = .45).

Hence, the predictive performance of all six models still substantially exceeded the predictive accuracy of the reference models.

Regarding the ordinal regression models, using the strength of evidence as predictor again led to the best model performance even in cross-study predictions, with correlations between model predictions (from study 1) and self-reports (in study 2) of r_s = .72 (SoE), and r_s = .68 (WoE), and correlations between model post-dictions (from study 2) and self-reports (in study 1) of r_s = .68 (SoE), and r_s = .65 (WoE). Moreover, also in terms of the accuracy, the model using the strength of evidence as predictor led to the best model performance when post-dicting from study 2 to study 1. However, when predicting from study 1 to study 2, the accuracies of the strength of evidence and the weight of evidence models were virtually identical (see Table 3). Finally, these two models clearly outperformed the reference models.

In the cross-study analyses, the proportion of identical predictions between these two models was slightly higher as compared to the within study analyses (see Table S3), yet still not at the upper bound (i.e., where each correct prediction of the worse model aligns with those of the better model) and thus still distinct in several cases (see also Figure S5). This is again highlighted in the larger number of distinct predictions made by the model with the strength of evidence as predictor, as opposed to the one with the weight of evidence as predictor (see Table 3).

Discussion

The cross-study analyses corroborated the conclusions drawn in study 1; namely, that the aspects' strength of evidence is the most important property of evidence for predicting self-reported risk preferences. As such, these analyses permitted ruling out a potential methodological confound due to the close temporal proximity between the section during which participants listed their aspects, and the section in which they self-reported their risk preferences.

Of note, although self-reported risk preferences showed a very high test–retest reliability across the two studies, some degree of intraindividual variability occurred. In light of this observation, some drop in model performance is naturally to be expected when making cross-study pre- and post-dictions. Notwithstanding this, and crucially, the models using the properties of evidence as predictors clearly outperformed the three reference models.

General Discussion

In the two studies presented in this article, we aimed to shed light on the information-integration processes underlying people's self-reports of their risk preferences, and to examine people's cognitive representations thereof. To this end, we made use of the process-tracing method of aspect listing and employed cognitive modeling to examine the extent to which different properties of evidence of the retrieved aspects are predictive of people's self-reports. Moreover, we investigated the stability of the "cognitive input" supposedly underlying people's self-reports (i.e., aspect- and evidence stability), the stability of the output (i.e., self-reported risk preferences), as well as the relation between stability in input and output. The results suggest three main take-home messages.

First, the two studies provide evidence for the internal validity of people's self-reports of their risk preferences. The desirable psychometric properties of the respective measures have increasingly been documented in recent research (e.g., Frey et al., 2017; Frey, Richter, et al., 2020; Mata et al., 2018), and the current analyses suggest a set of reasons for these observations. Specifically, people's self-reports ap-

pear to be the systematic result of a quantifiable information-integration process (see also Jarecki & Wilke, 2018): The aspects that participants retrieved from their memory during this process proved to be highly predictive for their self-reports—within and across the two studies reported here. Moreover, the aspects that form the input to this judgment-formation process mostly comprise situations that people frequently experience in their daily lives (see also Arslan et al., 2020; Schimmack et al., 2002; van der Linden, 2014; Weber, 2006)—rather than rare and exceptional, and thus potentially less diagnostic experiences.

Second, our model comparison unveiled several quantitative and qualitative properties of this information-integration process. From a theoretical point of view, people may consider three different properties of the retrieved information, namely, the weight, strength, and order of evidence. Whereas some research has primarily explored the weight of evidence of retrieved information (i.e., "how many pieces of information support a particular judgment?"; Jarecki & Wilke, 2018), here we also took into account the role of the other two dimensions. Our results indicated that the order of evidence may be largely irrelevant in this context, and people appeared to be particularly sensitive to the strength of evidence of retrieved information; that is, *how strongly* different aspects support a particular judgment concerning their risk preferences. This observation resonates with findings from other domains of judgment and decision making (Griffin & Tversky, 1992; Kvam & Pleskac, 2016) and suggests that similar information-integration processes may operate in judgment formation based on internal and external samples.

Third, our longitudinal analyses across the two studies illustrated that the properties of the cognitive input in people's judgments remained considerably stable (i.e., evidence stability), thus providing an explanation for why self-report measures of risk preference may show a high test–retest reliability (i.e., substantially higher than behavioral measures of the same construct; Frey et al., 2017; Lönnqvist et al., 2015; Mata et al., 2018). Specifically, the extent to which the strength of evidence of participants' listed aspects changed across time was strongly associated with changes in their self-reported risk preferences. The process of rendering self-reports arguably involves drawing internal samples of idiosyncratic experiences and past behaviors. According to our analyses, people retrieve aspects that are quite diverse in terms of their specific content, but highly similar in terms of their strength of evidence—and it was the latter dimension that people were mostly sensitive to when rendering a self-report. The high degree of evidence stability suggests that the retrieved experiences, albeit diverse, tend to support a similar degree of risk seeking or risk avoidance. In short, when rendering self-reports people may internally aggregate over different situations, and because the resulting self-reports thus encompass diverse settings, they may end up being predictive for a wide range of future behaviors and outcomes (e.g., Duckworth, Gendler, & Gross, 2016; Duckworth & Yeager, 2015). This interpretation likely extends beyond self-reports of risk preference to domain-specific conceptions of risk preferences, and may also apply in other areas of psychological research (e.g., Blais & Weber, 2006; Duckworth & Kern, 2011; Eisenberg et al., 2019; Jarecki & Wilke, 2018; Sharma, Markon, & Clark, 2014; Wilke et al., 2014).

Cognitive Modeling as a Tool to Unpack Self-Reports?

As the three take-home messages above illustrate, we believe that our approach of using a process-tracing method—along with cognitive modeling—was highly instrumental in uncovering the information-integration processes and cognitive representations underlying people's self-reports. This approach rests on the assumption that judgment and decision making typically involve information-sampling and -integration processes, with information being sampled from either internal or external sources (Fiedler & Juslin, 2005; E. J. Johnson et al., 2007; Juslin & Olsson, 1997). Yet, to what extent can one be confident of having identified the true underlying process? Clearly, the various models implemented here remain approximations of the true psychological processes that may operate in people's minds, and even good model predictions do not guarantee that one has identified the "correct" process (see Roberts & Pashler, 2000). Hence, by increasing the degree of observable data beyond the self-reported aspects we used as input in our approach (e.g., reaction times, physiological indicators), lower-level and more fine-grained inferences concerning specific cognitive processes will become possible.

Nevertheless, we believe that the clear systematicity with which aspects and self-reported risk preference were related (within and across studies), the pattern with which stability in the aspects' strength of evidence was associated with stability in self-reported risk preference, and finally, the strong agreement in the strength of evidence as indicated by participants and by external raters are all indicators for the robustness of the approach implemented here. That said, in what follows we would like to discuss potential limitations of our studies and suggest avenues for further research in the future.

Limitations and Further Research

Aspect listing. One potential issue of aspect listing—at least when implemented in the traditional way (i.e., within one session only)—consists of the close temporal proximity between the listing of aspects and providing the self-report itself, hence potentially inflating the respective consistency. Our design with a retest study permitted addressing this issue directly: Even in the cross-study analyses the predictive accuracies of the various models were high and far superior compared to those of sociodemographic predictors. This

suggests that the good performance of the cognitive models does not merely reflect a methodological artifact.

Yet, there are potentially even more fundamental issues related to the method of aspect listing that are worthy of a careful discussion. As outlined in our introduction, a basic motivation for employing aspect listing is to avoid having to prompt respondents to engage in introspection *in hindsight*; that is, to reflect on how they had rendered a previous self-report. Specifically, it has been argued that such retrospective metacognitive judgments may be unreliable, as people lack sufficient insight into the cognitive processes underlying their own judgments (Nisbett & Bellows, 1977; Nisbett & Wilson, 1977). Thus, to avoid this potential issue, methods such as aspect listing or think-aloud protocols aim to trace information processing on the fly (e.g., Ericsson & Simon, 1980, 1993). Naturally, there are also some intricacies with this approach, as it evidently rests on the assumption that people are capable of providing veridical reports of their own, ongoing thoughts—and this assumption may not always be met, at least not entirely: On the one hand, the task of sequentially typing in one's ongoing thoughts may alter the judgment-formation process to be more systematic, thus potentially leading to a more structured way of rendering a self-report (see Ericsson & Simon, 1980; Fox, Ericsson, & Best, 2011). To illustrate, the somewhat stronger bimodal distribution of participants' self-reports in our studies (i.e., as compared to in previous studies; e.g., Dohmen et al., 2011; Frey et al., 2017) might be a manifestation of this possibility—although there were no indications for systematic mean differences, depending on whether self-reports were provided after or before the actual aspect listing (as investigated in a pilot study, see methods section of study 1). On the other hand, assuming that the method of aspect listing does not overly distort the ongoing judgment-formation process, one still cannot be entirely sure that the listed aspects reflect fully accurate memories, as memories of everyday life events could be altered and transformed (for reviews, see Koriat, 2007; Koriat et al., 2000). Thus, in future research it will be useful to test whether our findings also hold for other process-tracing methods such as think-aloud protocols, which might be more robust in this regard (Fox et al., 2011). Relatedly, it may be worthwhile to test the extent to which particular contexts trigger the retrieval of specific (classes of) aspects, which could in principle explain why aspect stability (but not evidence stability) was low across the two studies conducted here. Taken together, people may not always have direct introspective access to the processes involved in their judgments and decisions (i.e., particularly when being prompted to reflect on such processes explicitly and in hindsight). Yet, under certain conditions and when using the appropriate methods they may indeed be able to report on their current thoughts quite accurately, thus providing reliable insight concerning the underlying cognitive processes (e.g., Adair & Spinner, 1981; Berger, Dennehy, Bargh, & Morsella, 2016; Ericsson & Simon, 1980, 1993; Hurlburt & Heavey, 2001; White, 1980).

Modeling approach. We have sampled diverse models from the literature on judgment and decision making that describe manifold information-integration processes, and which cover a wide space ranging from simple heuristics to learning models. As the empirical data imposed some constraints concerning the level of detail with which fine-grained model comparisons were possible, we additionally relied on a more general model comparison—focusing on the distinction between the strength of evidence and the weight of evidence. In the future, the employment of yet other process-tracing approaches (see points discussed above) might allow for more fine-grained analyses in this respect.

Moreover, as we modeled one self-report per participant, we estimated the free parameters across individuals. Although this is a widespread procedure in various applications of cognitive modeling (e.g., Birnbaum, 2008; Erev, Ert, Plonsky, Cohen, & Cohen, 2017; Erev et al., 2010), this approach is not without its problems. For example, not all participants may rely on the same information-integration processes (e.g., Frey, Rieskamp, & Hertwig, 2015; Mata, von Helversen, & Rieskamp, 2010; Payne et al., 1988) and/or may be best described with the same parameter values (e.g., Kellen et al., 2016; Pedroni et al., 2017). In short, it is unclear to what extent findings based on the *interindividual* level generalize to the *intraindividual* level (Molenaar, 2004; Molenaar & Campbell, 2009), and future research is thus needed to clarify a potential heterogeneity between different persons' cognitive processes.

Outcome measures. Finally, future research may also investigate to what extent our findings extend to self-reports of domain-specific risk preferences. Jarecki and Wilke (2018) have examined how cognitive processes potentially vary across different (evolutionary) content-domains (see also Wilke et al., 2014). Similar analyses, yet including models that take into account the strength of evidence of retrieved information, could thus also be conducted for domain-specific risk preferences as are often assessed in psychological research (Rolison & Shenton, 2020; Weber et al., 2002). One may expect that aspects retrieved for specific domains of life (e.g., recreation, health, finance) may be more heterogeneous across domains, but more homogeneous within, as compared to those retrieved in response to a domain-general question as investigated here—which may ultimately increase aspect stability.

Conclusions

Zooming out, our approach to modeling people's self-reported risk preferences involves several contributions that inform psychological assessment in general, and provides

theoretical and measurement-related insight into the construct of risk preference more specifically.

First, we bridged two methodological approaches that are too often employed separately; that is, we investigated self-reported preferences (as typically employed in psychometric research relying on questionnaires) by implementing cognitive modeling using a range of different models. Integrating these approaches proved helpful for a better understanding of the construct validity of self-reported risk preference, and we hope that our approach will inspire similar applications in other areas of psychological research in the future.

Second, our investigations provide substantial evidence that self-reports of risk preference are robustly rooted in people's idiosyncratic experiences, and are thus internally valid. Specifically, the desirable psychometric properties of respective self-report measures—here tapping risk preference, but potentially also in the case of self-reports of other psychological constructs—may emerge as the result of an information-integration process that aggregates multiple samples that people draw from their autobiographical memory.

Third and finally, our findings have an important implication for applied settings: Risk preferences can have a dramatic impact on important life outcomes, and are thus frequently assessed in various real-life contexts, such as concerning health- and safety-related matters or when designing investment portfolios. In doing so, the tool of choice is often one of numerous self-report measures. These measures may be not only frugal in their application—but according to our findings also sound from a psychological perspective.

Author Contributions

All authors developed the study concepts and contributed to the design. M.D.S. and F.I.S performed the data collection, analysis, and interpretation under the supervision of R.F. M.D.S. and R.F. wrote the manuscript, and F.I.S. provided significant revisions. All authors approved the final version of the article.

References

Adair, J. G., & Spinner, B. (1981). Subjects' Access to Cognitive Processes: Demand Characteristics and Verbal Report. *Journal for the Theory of Social Behaviour*, *11*(1), 31–52. doi: 10.1111/j.1468-5914.1981.tb00021.x

Andreoni, J., & Kuhn, M. A. (2019). Is it safe to measure risk preferences? Assessing the completeness, predictive validity, and measurement error of various techniques. *Working Paper*. Retrieved from https://static1.squarespace.com/static/5c79b3d29b8fe82f5cb96360/t/5cc0debb71c10bd5d9ab45f3/1556143804348/mCRB_WP.pdf

Appelt, K. C., Hardisty, D. J., & Weber, E. U. (2011). Asymmetric discounting of gains and losses: A query theory account. *Journal of Risk and Uncertainty*, *43*(2), 107–126. doi: 10.1007/s11166-011-9125-1

Arslan, R. C., Brümmer, M., Dohmen, T., Drewelies, J., Hertwig, R., & Wagner, G. G. (2020). How people know their risk preference. *Scientific Reports*, *10*(1), 15365. doi: 10.1038/s41598-020-72077-5

Balatel, A., Boero, R., Jonaityte, I., Monti, M., Novarese, M., & Pacelli, V. (2013). Beyond the MiFID: Envisioning cognitively suitable and representationally supportive approaches to assessing investment preferences for more informed financial decisions. *CAREFIN Occasional Paper*. Retrieved from http://hdl.handle.net/11858/00-001M-0000-0024-ED29-9

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. doi: 10.1016/j.jml.2012.11.001

Barron, G., & Yechiam, E. (2009). The coexistence of overestimation and underweighting of rare events and the contingent recency effect. *Judgment and Decision Making*, *4*, 447–460. Retrieved from http://journal.sjdm.org/9729b/jdm9729b.pdf

Beauchamp, J., Cesarini, D., & Johannesson, M. (2017). The psychometric and empirical properties of measures of risk preferences. *Journal of Risk and Uncertainty*, *54*(3), 203-237. doi: 10.1007/s11166-017-9261-3

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research*, *30*(5), 961–981. doi: 10.1287/opre.30.5.961

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 183–200. doi: 10.1037/h0024835

Berg, J., Dickhaut, J., & McCabe, K. (2005). Risk preference instability across institutions: A dilemma. *Proceedings of the National Academy of Sciences*, *102*(11), 4209-4214. doi: 10.1073/pnas.0500333102

Berger, C. C., Dennehy, T. C., Bargh, J. A., & Morsella, E. (2016). Nisbett and Wilson (1977) Revisited: The Little That We Can Know and Can Tell. *Social Cognition*, *34*(3), 167–195. doi: 10.1521/soco.2016.34.3.167

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*(2), 463 - 501. doi: 10.1037/0033-295X.115.2.463

Blais, A.-R., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, *1*, 33–47. doi: 10.1037/t13084-000

Broomell, S. B., Budescu, D. V., & Por, H.-H. (2011). Pairwise comparisons of multiple models. *Judgment and Decision Making*, *6*(8), 821–831. Retrieved from http://journal.sjdm.org/11/m09/m09.html

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. doi: 10.1177/1745691610393980

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*(3), 367–383. doi: 10.1037/0033-2909.125.3.367

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via amazon mturk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*, 2156 - 2160. doi: 10.1016/j.chb.2013.05.009

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475-494. doi: 10.1177/001316446600600405

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. doi: 10.1037/0003-066X.34.7.571

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*(2), 95–106. doi: 10.1037/h0037613

Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, *95*(3), 542-575. doi: 10.1037/0033-2909.95.3.542

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

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2016). Situational strategies for self-control. *Perspectives on Psychological Science*, *11*(1), 35-55. doi: 10.1177/1745691615623247

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*, 259-268. doi: 10.1016/j.jrp.2011.02.004

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087-1101. doi: 10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*, 237-251. doi: 10.3102/0013189X15584327

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, *10*, 2319. doi: 10.1038/s41467-019-10301-1

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, *124*(4), 369–409. doi: 10.1037/rev0000062

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., … Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*(1), 15–47. doi: 10.1002/bdm.683

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. doi: 10.1037/0033-295X.87.3.215

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT press.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. doi: 10.3758/BRM.41.4.1149

Ferrarini, G., & Wymeersch, E. (2006). *Investor protection in Europe: Corporate law making, the MiFID and beyond*. Oxford, UK: Oxford University Press.

Fiedler, K., & Juslin, P. (2005). *Information sampling and adaptive cognition*. New York: Cambridge University Press.

Fiedler, K., Renn, S.-Y., & Kareev, Y. (2010). Mood and judgments based on sequential sampling. *Journal of Behavioral Decision Making*, *23*(5), 483–495. doi: 10.1002/bdm.669

Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., & Combs, B. (1978). How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sciences*, *9*(2), 127–152. doi: 10.1007/BF00143739

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316–344. doi: 10.1037/a0021663

Frey, R., Duncan, S., & Weber, E. U. (2020). Towards a typology of risk preference: Four risk profiles describe two thirds of individuals in a large sample of the U.S. population. *PsyArXiv Preprint*. doi: 10.31234/osf.io/yjwr9

Frey, R., Mata, R., & Hertwig, R. (2015). The role of cognitive abilities in decisions from experience: Age differences emerge as a function of choice set size. *Cognition*, *142*, 60–80. doi: 10.1016/j.cognition.2015.05.004

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, *3*, e1701381. doi: 10.1126/sciadv.1701381

Frey, R., Richter, D., Schupp, J., Hertwig, R., & Mata, R. (2020). Identifying robust correlates of risk preference: A systematic approach using specification curve analysis. *Journal of Personality and Social Psychology*. doi: 10.1037/pspp0000287

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

Galizzi, M. M., Machado, S. R., & Miniaci, R. (2016). Temporal stability, cross-validity, and external validity of risk preferences measures: Experimental evidence from a UK representative sample. *London School for Economics and Political Science Working Paper*. doi: 10.2139/ssrn.2822613

Galton, F. (1874). *English men of science*. London: Macmillan.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143. doi: 10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. doi: 10.1037/0033-295X.103.4.650

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In G. Gigeren-

zer, P. M. Todd, & The ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 75–95). New York: Oxford University Press.

Gluth, S., & Jarecki, J. B. (2019). On the Importance of Power Analyses for Cognitive Modeling. *Computational Brain & Behavior*, *2*(3-4), 266–270. doi: 10.1007/s42113-019-00039-w

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2018). *rstanarm: Bayesian applied regression modeling via Stan.* Retrieved from http://mc-stan.org/ (R package version 2.17.4)

Greifeneder, R., & Schwarz, N. (2014). Metacognitive processes and subjective experiences. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 314–327). New York: The Guilford Press.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435. doi: 10.1016/0010-0285(92)90013-R

Guttman, L. (1944). A Basis for Scaling Qualitative Data. *American Sociological Review*, *9*(2). doi: 10.2307/2086306

Hammond, K. R., & Stewart, T. R. (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.

Hardisty, D. J., Johnson, E. J., & Weber, E. U. (2010). A dirty word or a dirty world? attribute framing, political affiliation, and query theory. *Psychological Science*, *21*, 86–92. doi: 10.1177/0956797609355572

Hastie, R., & Dawes, R. M. (2001). A general framework for judgment. In *Rational choice in an uncertain world: The psychology of judgment and decision making* (pp. 47–69). Thousand Oaks, CA: Sage Publications, Inc.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review*, *93*(3), 258–268. doi: 10.1037/0033-295X.93.3.258

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539. doi: 10.1111/j.0956-7976.2004.00715.x

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2006). The role of information sampling in risky choice. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 72–91). New York: Cambridge University Press.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*(2), 225–237. doi: 10.1016/j.cognition.2009.12.009

Highhouse, S., & Gallo, A. (1997). Order effects in personnel decision making. *Human Performance*, *10*, 31-46. doi: 10.1207/s15327043hup1001_2

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1-55. doi: 10.1016/0010-0285(92)90002-J

Hurlburt, R. T., & Heavey, C. L. (2001). Telling what we know: Describing inner experience. *Trends in Cognitive Sciences*, *5*(9), 400–403. doi: 10.1016/S1364-6613(00)01724-1

Jarecki, J. B., & Wilke, A. (2018). Into the black box: Tracing information about risks related to 10 evolutionary problems. *Evolutionary Behavioral Sciences*, *12*, 230–244. doi: 10.1037/ebs0000123

Jobe, J. B. (2000). Cognitive processes in self-report. In A. A. Stone, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 25–29). Mahwah, NJ: Lawrence Erlbaum.

Jobe, J. B. (2003). Cognitive psychology and self-reports: Models and methods. *Quality of Life Research*, *12*, 219–227. doi: 10.1023/A:1023279029852

Jobe, J. B., & Herrmann, D. J. (1996). Implications of models of survey cognition for memory theory. In D. J. Herrman, C. McEvoy, C. Hertzog, P. Hertel, & M. K. Johnson (Eds.), *Basic and applied memory research* (pp. 193–205). New York: Laurence Erlbaum Associates, Inc.

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 461–474. doi: 10.1037/0278-7393.33.3.461

Johnson, J. G., & Raab, M. (2003). Take the first: Option-generation and resulting choices. *Organizational Behavior and Human Decision Processes*, *91*, 215–229. doi: 10.1016/S0749-5978(03)00027-X

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, *104*, 344-366. doi: 10.1037/0033-295X.104.2.344

Kahneman, D., & Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. *Proceedings of the National Academy of Sciences*, *107*(38), 16489–16493. doi: 10.1073/pnas.1011492107

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. doi: 10.2307/1914185

Kellen, D., Pachur, T., & Hertwig, R. (2016). How (in)variant are subjective representations of described and experienced risk and rewards? *Cognition*, *157*, 126–138. doi: 10.1016/j.cognition.2016.08.020

Kendall, M. G. (1948). *Rank correlation methods*. London, UK: Griffin.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*(4), 609–639. doi: 10.1037/0033-295X.100.4.609

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The cambridge handbook of consciousness* (pp. 289–325). New York: Cambridge University Press.

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology*, *51*(1), 481–537. doi: 10.1146/annurev.psych.51.1.481

Kvam, P. D., & Pleskac, T. J. (2016). Strength and weight: The determinants of choice and confidence. *Cognition*, *152*, 170–180. doi: 10.1016/j.cognition.2016.04.008

Lejarraga, T., Frey, R., Schnitzlein, D. D., & Hertwig, R. (2019). No effect of birth order on adult risk taking. *Proceedings of the National Academy of Sciences*, *116*, 6019-6024. doi: 10.1073/pnas.1814153116

Lerner, J. S., Gonzalez, R. M., Small, D. A., & Fischhoff, B. (2003). Effects of fear and anger on perceived risks of terrorism: A national field experiment. *Psychological Science*, *14*(2),

144–150. doi: 10.1111/1467-9280.01433

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22 140*, 55-55.

Lindskog, M., Winman, A., & Juslin, P. (2013). Naïve point estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 782-800. doi: 10.1037/a0029670

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*(2), 267-286. doi: 10.1037/0033-2909.127.2.267

Lönnqvist, J.-E., Verkasalo, M., Walkowitz, G., & Wichardt, P. C. (2015). Measuring individual risk attitudes in the lab: Task or ask? An empirical comparison. *Journal of Economic Behavior & Organization*, *119*, 254-266. doi: 10.1016/j.jebo.2015.08.003

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*(368), 805–824. doi: 10.2307/2232669

Lopes, L. L. (1984). Risk and distributional inequality. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 465-485. doi: 10.1037/0096-1523.10.4.465

Maccrimmon, K. R., & Wehrung, D. A. (1985). A portfolio of risk measures. *Theory and Decision*, *19*(1), 1–29. doi: 10.1007/BF00134352

Mamerow, L., Frey, R., & Mata, R. (2016). Risk taking across the life span: A comparison of self-report and behavioral measures of risk taking. *Psychology and Aging*, *31*, 711–723. doi: 10.1037/pag0000124

Mata, R., Frey, R., Richter, D., Schupp, J., & Hertwig, R. (2018). Risk preference: A view from psychology. *Journal of Economic Perspectives*, *32*(2), 155–172. doi: 10.1257/jep.32.2.155

Mata, R., Josef, A. K., & Hertwig, R. (2016). Propensity for risk taking across the life span and around the globe. *Psychological Science*, *27*(2), 231–243. doi: 10.1177/0956797615617811

Mata, R., von Helversen, B., & Rieskamp, J. (2010). Learning to choose: Cognitive aging and strategy selection learning in decision making. *Psychology and Aging*, *25*(2), 299. doi: 10.1037/a0018923

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81-90. doi: 10.1037/0022-3514.52.1.81

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*(6), 423–429. doi: 10.1111/j.1467-9280.1997.tb00455.x

Mishra, S., Barclay, P., & Sparks, A. (2017). The relative state model: Integrating need-based and ability-based pathways to risk-taking. *Personality and Social Psychology Review*, *21*(2), 176-198. doi: 10.1177/1088868316644094

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. doi: 10.1207/s15366359mea0204_1

Molenaar, P. C., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*(2), 112–117. doi: 10.1111/j.1467-8721.2009.01619.x

Nicholson, N., Soane, E., Fenton-O'Creevy, M., & Willman, P. (2005). Personality and domain-specific risk taking. *Journal of Risk Research*, *8*(2), 157–176. doi: 10.1080/1366987032000123856

Nisbett, R. E., & Bellows, N. (1977). Verbal reports about causal influences on social judgments: Private access versus public theories. *Journal of Personality*, *35*(9), 613–624. doi: 10.1037/0022-3514.35.9.613

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. doi: 10.1037/0033-295X.84.3.231

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419. doi: 10/10630a/jdm10630a

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 534–552. doi: 10.1037/0278-7393.14.3.534

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, *1*, 803–809. doi: 10.1038/s41562-017-0219-x

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior Research Methods*, *46*(4), 1023–1031. doi: 10.3758/s13428-013-0434-y

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358-367. doi: 10.1037/0033-295X.107.2.358

Rohrer, J. M. (2017). Test–retest reliabilities of scales included in the socio-economic panel study. *PsychArxiv Preprint*. doi: 10.31219/osf.io/3ncbt

Rolison, J. J., & Shenton, J. (2020). How much risk can you stomach? Individual differences in the tolerance of perceived risk across gender and risk domain. *Journal of Behavioral Decision Making*, *33*, 63–85. doi: 10.1002/bdm.2144

Schimmack, U., Diener, E., & Oishi, S. (2002). Life-satisfaction is a momentary judgment and a stable personality characteristic: The use of chronically accessible and stable sources. *Journal of Personality*, *70*(3), 345-384. doi: 10.1111/1467-6494.05008

Schulte-Mecklenbeck, M., Johnson, J. G., Böckenholt, U., Goldstein, D. G., Russo, J. E., Sullivan, N. J., & Willemsen, M. C. (2017). Process-tracing methods in decision making: On growing up in the 70s. *Current Directions in Psychological Science*, *26*(5), 442-450. doi: 10.1177/0963721417708229

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *The American Journal of Evaluation*, *22*(2), 127-160. doi: 10.1016/S1098-2140(01)00133-3

Sharma, L., Markon, K. E., & Clark, L. A. (2014). Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bul-*

letin, *140*(2), 374-408. doi: 10.1037/a0034418

Sitkin, S. B., & Pablo, A. L. (1992). Reconceptualizing the determinants of risk behavior. *Academy of Management Review*, *17*, 9–38. doi: 10.5465/amr.1992.4279564

Steiner, M. D., & Frey, R. (2021). Representative design in psychological assessment: A case study using the balloon analogue risk task (BART). *Journal of Experimental Psychology: General*, Advance Online Publication. doi: 10.1037/xge0001036

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT press.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273-286. doi: 10.1037/h0070288

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*(4), 529-554. doi: 10.1086/214483

Tisdall, L., Frey, R., Horn, A., Ostwald, D., Horvath, L., Blankenburg, F., ... Mata, R. (2020). Brain-behavior associations for risk taking depend on the measures used to capture individual differences. *Frontiers in Behavioral Neuroscience*, *14*. doi: 10.3389/fnbeh.2020.587152

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, *315*(5811), 515–518. doi: 10.1126/science.1134239

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.

van der Linden, S. (2014). On the relationship between personal experience, affect and risk perception: The case of climate change. *European Journal of Social Psychology*, *44*, 430–440. doi: 10.1002/ejsp.2008

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4

Wang, X. T., Kruger, D. J., & Wilke, A. (2009). Life history variables and risk-taking propensity. *Evolution and Human Behavior*, *30*(2), 77–84. doi: 10.1016/j.evolhumbehav.2008.09.006

Weber, E. U. (2006). Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). *Climatic Change*, *77*, 103–120. doi: 10.1007/s10584-006-9060-3

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., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., & Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice: A query-theory account. *Psychological Science*, *18*, 516–523. doi: 10.1111/j.1467-9280.2007.01932.x

Weber, E. U., & Milliman, R. A. (1997). Perceived risk attitudes: Relating risk perception to risky choice. *Management Science*, *43*(2), 123–144. doi: 10.1287/mnsc.43.2.123

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review*, *87*(1), 105–112.

Wilke, A., Sherman, A., Curdt, B., Mondal, S., Fitzgerald, C., & Kruger, D. J. (2014). An evolutionary domain-specific risk scale. *Evolutionary Behavioral Sciences*, *8*(3), 123–141. doi: 10.1037/ebs0000011

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. doi: 10.1177/1745691617693393

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review*, *12*, 387–402. doi: 10.3758/BF03193783

Zuckerman, M. (2002). *Sensation Seeking and Risky Behavior*. Binghamton, NY: Maple-Vail Press.

Table 1

Assessment of the Aspects' Properties in Study 1

Property/Source	Derived From	Operationalization
Properties used to model self-reported risk preferences		
Strength of evidence	Rating by participant	"How strongly does your description above support that you seek risks vs. that you avoid risks?" (ranging from 50, labeled "strong support for risk seeking," to -50, labeled "strong support for risk avoidance")
Weight of evidence	Binarized strength of evidence	Number of pro-aspects (strength of evidence > 0) and number of contra-aspects (strength of evidence < 0). Neutral aspects (strength of evidence = 0) were ignored in the analysis.
Order of evidence	Sequence of the listed aspects	Depending on the models implementing serial-position effects (see QT and VUM).
Properties to explore sources and content of evidence		
Personal experience	Rating by participant	"Does your description above include a personal experience?"
Social comparison	Rating by participant	"Does your description above include a comparison with another person?"
Frequency in daily life	Rating by participant	"How frequently do you normally experience or do what you described above?"
Active choice vs. passive experience	Rating by participant	"Does what you described above include something you actively chose or something you passively experienced?"
Controllability	Rating by participant	"Are the outcomes and consequences of what you described above controllable or uncontrollable for you?"
Sentiment	Wording of aspects	A score of the sentiment (positive or negative) of an aspect.

Note: The ratings concerning personal experience, social comparison, active choice, and controllability were binary ("Yes, it includes a personal experience" vs. "No, it does not include a personal experience" for the personal experience rating; "Yes, it includes a comparison with another person" vs. "No, it does not include a comparison with another person" for the rating of social comparison; "Rather controllable" vs. "Rather uncontrollable" for the rating of controllability; and "Active choice" vs. "Passive experience" for the rating concerning active choice vs. passive experience). Frequency was assessed categorically, with the categories "Once per day", "Once per week", "Once per month", "Once per year", "Less than once per year, but on a regular basis", and "Less than once per year, only once or a few times so far". In all ratings on the sources and content, a "Not applicable" option could be selected.

Table 2

Performance of the Different Ordinal Regression Models.

Model	Accuracy	κ	Distinct Predictions	ELPD
Study 1				
SoE	.42	.37	7	0 [0, 0]
WoE	.34	.28	5	-57.4 [-75.0, -39.8]
5 soc. dem. pred.	.24	.16	3	-194.6 [-219.6, -169.6]
Sex	.19	.11	2	-194.7 [-218.3, -171.1]
Age	.20	.12	3	-195.8 [-220.0, -171.6]
Study 2				
SoE	.33	.27	7	0 [0, 0]
WoE	.31	.24	5	-21.4 [-34.4, -8.4]
5 soc. dem. pred.	.23	.16	4	-84.8 [-111.6, -58.6]
Sex	.20	.12	3	-82.4 [-109.4, -55.4]
Age	.21	.13	1	-82.6 [-110.2, -55.0]

Note: Results are based on ordinal regression models (not cross-validated). Accuracy = The proportion of correctly predicted categories (ratings between 0 and 10). κ = Cohen's kappa, with a chance level of 1/11. Distinct Predictions = The number of distinct/unique predictions made by a model (all numbers from 0 to 10 occurred in the empirical data). ELPD = Estimate of the leave-one-out information criterion based expected log predictive density for a new dataset, relative to the best model (i.e., SoE)—where lower numbers indicate worse model fit. ± 2 standard errors interval are given in brackets. SoE = Mean strength of evidence per participant as predictor. WoE = Mean weight of evidence per participant as predictor. 5 soc. dem. pred. = Age, sex, years of education, income, and employment status as predictors. Sex = Sex as predictor. Age = Age as predictor.

Table 3

Performance of the Different Ordinal Regression Models in the Cross-Study Analyses.

Model	Accuracy	κ	Distinct Predictions
Fitting in study 1, prediction to study 2			
SoE	.28	.21	7
WoE	.29	.22	5
5 soc. dem. pred.	.26	.19	3
Sex	.21	.13	2
Age	.21	.13	3
Fitting in study 2, post-diction to study 1			
SoE	.35	.29	7
WoE	.31	.25	6
5 soc. dem. pred.	.24	.16	3
Sex	.22	.14	3
Age	.21	.14	1

Note: Predictions are based on ordinal regression models fit within study 1 and 2 (shown in Table 2). Accuracy = The proportion of correctly predicted categories (ratings between 0 and 10). κ = Cohen's kappa, with a chance level of 1/11. Distinct Predictions = The number of distinct/unique predictions made by a model (all numbers from 0 to 10 occurred in the empirical data). SoE = Mean strength of evidence per participant as predictor. WoE = Mean weight of evidence per participant as predictor. 5 soc. dem. pred. = Age, sex, years of education, income, and employment status as predictors. Sex = Sex as predictor. Age = Age as predictor.