How people know their risk preference

Ruben C. Arslan[*1], Martin Brümmer[2], Thomas Dohmen[3,4,5,6,7], Johanna Drewelies[8], Ralph Hertwig[†1], Gert G. Wagner[†1,4,6]

* corresponding author

† RH and GGW contributed equally.

Abstract

People differ in their willingness to take risks. Recent work found that a dominant class of measures, revealed preference tasks (e.g., laboratory lotteries), appear not to tap into stable individual differences, whereas survey-based stated preferences are stable and predict real-world risk taking across different domains. How can stated preferences, often criticised as inconsequential ("cheap talk"), be more valid and predictive than controlled, incentivized lotteries? In our multi-method study, over 3,000 respondents from population samples answered a single widely used and predictive risk preference question. Respondents then explained the reasoning behind their answer. They tended to recount diagnostic behaviours and experiences, focusing on voluntary, consequential acts and experiences from which they seemed to infer their risk preference. We found that third-party readers of respondents' brief memories and explanations reached similar inferences about respondents' preferences, indicating the intersubjective validity of this information. Our results also shed light on the process of preference formation through experience over the lifespan. Finally, stated risk preferences may capture preferences revealed in behaviours in the wild better than the contrived behavioural tasks preferred in economics because they permit people to draw upon their own understanding of what constitutes diagnostic behaviours and experiences.

Keywords: risk preferences, self-report, self-perception

Affiliations:

1. Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin
2. University of Leipzig
3. Institute for Applied Microeconomics, University of Bonn
4. Institute of Labor Economics (IZA), Bonn, Germany
5. Maastricht University
6. German Institute for Economic Research
7. CESifo GmbH, Munich
8. Humboldt University Berlin

Corresponding author: Ruben C. Arslan,
Center for Adaptive Rationality, Max Planck Institute for Human Development,
Lentzeallee 94, 14195 Berlin, Germany. E-mail: ruben.arslan@gmail.com

Introduction

Consequential decisions about health, finances, and relationships often invoke the question of how much risk one is willing to take. Unsurprisingly then, risk preferences are widely studied in experimental economics; personality, cognitive, and clinical psychology; and even animal personality research.[1–4] Measures of risk preference can help predict a wide range of behaviours, from smoking and pathological gambling[5] to self-employment and holding stocks.[6–9]

Two very different measurement traditions have investigated risk preferences in humans. The *revealed preference* approach, common in economics, has sought to study choices under risk in the field[10] and in the laboratory.[11] The paradigmatic research designs in this tradition are observational studies of real behaviours (e.g., consumption and saving) and choices between monetary lotteries. At the same time, personality and clinical psychologists, as well as some economists, have used a *stated preference* approach. Here, people are asked to state their willingness to take risks, either using general questions or hypothetical scenarios. Our goal is to explain why and how stated preferences are informative by embedding them in the literature on self-perception and self-insight. In doing so we not only provide insight into the measurement of preferences, but also highlight the role that experiences play in preference formation over the lifespan.

Economists have been particularly skeptical about the validity of stated preferences, where (un)truthful answers have no material consequences (e.g. [12]); therefore, to verify the assumptions made to infer preferences from observations of real life behaviour, such as stability over time and the ability to control factors other than preferences, they have typically turned to revealed preference measures, which offered greater control over confounding factors while still measuring “real” behaviour (see [13–15]). Ironically, when researchers compared revealed and stated risk preference measures systematically,[5,16–18] they found that the behavioural measures used in the revealed preference approach consistently underperform relative to the stated preference measures in terms of reliability, stability, construct, and predictive validity.[4,13] Specifically, the former did not correlate across measures, meaning that they did not capture a latent preference that drives behaviour across different choice situations—even when differences between tasks were abstracted away by modelling the decision process.[19] In

contrast, the stated risk preferences correlated across items and questionnaires and suggested the existence of a general risk factor. Finally, convergence between revealed and stated preferences has been found to be low, particularly when confounders like age and gender are kept constant.[5,9,20,21]

Copious research has investigated the cognitive processes that underlie behaviour (e.g., choice) in the lab-based revealed preferences approach.[19,22] By contrast, little is known about the processes that shape responses in the stated preference approach (but see [23,24]). This gap may be a factor behind many economists not adopting the stated preference approach. Although self-reports are widely used in psychology, their accuracy is often contended, with some researchers emphasizing their context sensitivity and potential for bias and self-enhancement[25,26] and others arguing that self-reports are often valid under real-world conditions.[27–30]

While few believe that people can draw on absolute, internal values to objectively report their preferences or personality, there is a reason to believe people have a keen sense of their own relative standing on these dimensions. Bem[31] argued that people's self-perception co-opts the abilities used for social perception. That is, the same ability of instant recognition that allows a person to call someone a crazy bastard when they sprint across a busy street[32] can also be applied by a person to themself. Bem and other social psychologists focused on explaining how this co-opted adaptation causes lapses in self-judgment,[33] whereas recent work draws on the concept of self–other knowledge asymmetries to explain why people know themselves better than others do in some but not all areas.[29,30] Such asymmetries may also explain some of the discrepancy in validity between stated and revealed preference measures: People's preference to take risks can be "revealed" in their choices and actions, but the very same choice could be—depending on their psychological state, their current needs, and their overall abilities[34,35]—a risk taken willingly, an impulse regretted immediately, a last resort when cornered, or child's play for the highly skilled. Unlike the decision maker, external observers cannot easily access these internal states to infer the preferences from the observed behaviour.

To unpack the process of self-perception we asked how people translate their memories and intuitions into an answer to the question "How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?" on a scale from 0 to 10

("unwilling to take risks" to "fully prepared to take risks"). This single question, the General Risk Question (GRQ;[6] has been used in several large and widely analyzed surveys.[36–38] The GRQ is predictive of real-world risk taking[6] and is one of the best indicators of the general factor of risk preferences.[5] Many genetic loci linked to risk preferences in a genome-wide association study were identified through the use of similar single-item questions.[39]

Here, we took a descriptive approach that lets participants speak because systematically varying questions, examples, and reference frames[40–42] would already require deviations from the widely used GRQ. Instead, we asked people to explain how they answered the GRQ and which risks they thought about in order to illuminate how people infer their own risk preferences from their decisions, indecisions, and regrets. We were interested in three aspects of how people evaluate their risk preferences:

1. What kind of risks do people consider when they judge themselves? Are these concrete everyday risks with clear consequences, or small, cumulative risks with stochastic consequences? Which reference frames do people use? And do they mainly think about risks they took and considered worthwhile, or do risks they avoided or regretted taking feature too?
2. Do age and gender affect the risks people invoke and experience?
3. Can independent third parties agree on what people's experiences say about their preferences?

We collected stated risk preferences as part of two large, age-heterogeneous survey studies in Germany: the 2017 interim survey of the BASE-II study[43] and the 2017/2018 Innovation Sample of the German Socioeconomic Panel (SOEP).[44] Across both studies, 3,493 respondents answered the GRQ. After doing so, they were asked to explain their response in closed-form questions about the social and temporal reference frames they had had in mind, as well in free-text questions about the topics and events they had thought about. They then listed the biggest risks they took in the past year. BASE-II respondents were also asked if the risks they had taken had been worthwhile.

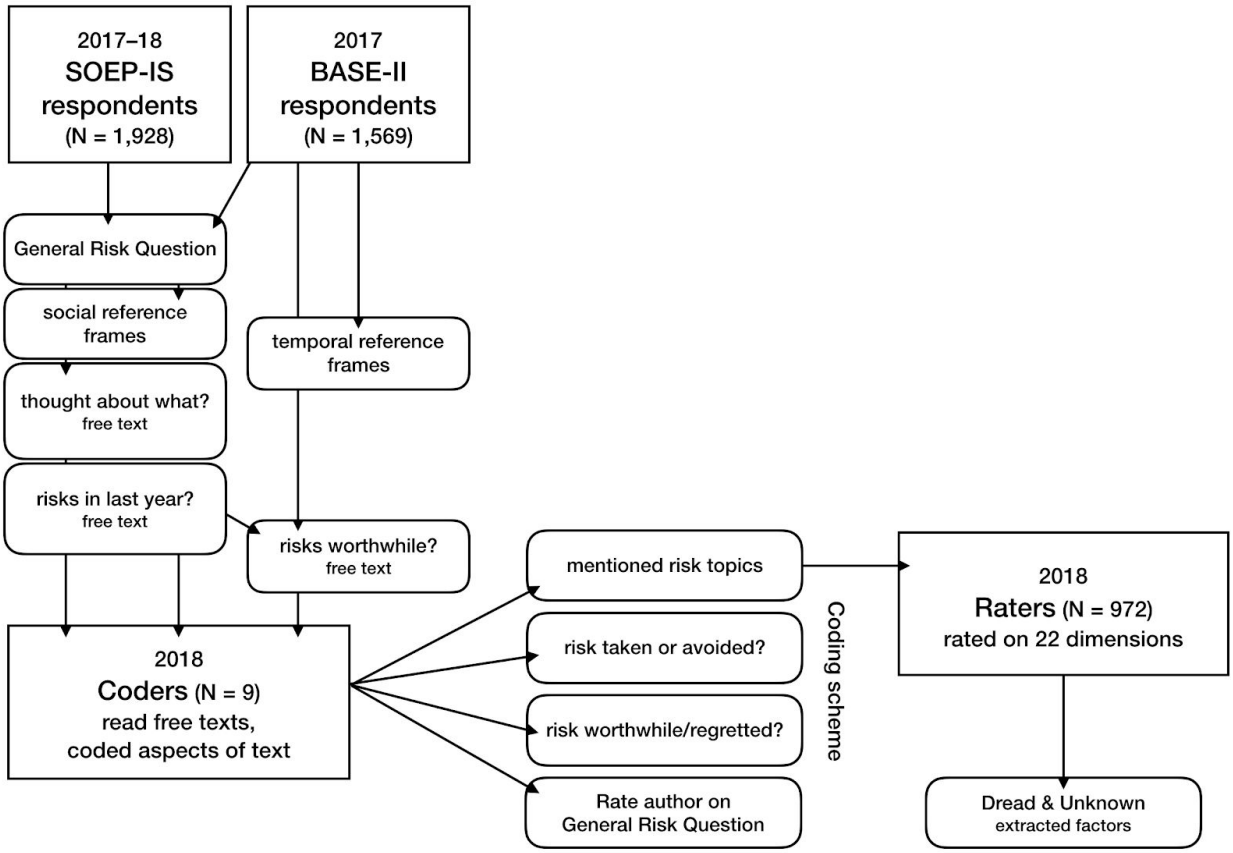

Figure 1: Flow chart of the data collection, coding, and rating steps. Boxes are samples, ellipses reflect steps in the data collecting and processing.

To quantify the topics featured in respondents' free-text answers, we conducted two further studies (Figure 1). We first designed a coding scheme with a list of risk factors. Participants of an online panel (n = 825) rated three to five specific risk factors or behaviours from this coding scheme, ranging from divorce to cycling, on 20 dimensions (e.g., voluntariness and immediacy) known in the literature[45,46] and on two dimensions to differentiate social from mortality risks. From 16 of these dimensions, we extracted the well-known factors *Dread* and *Unknown*[45] in a confirmatory factor analysis (see Supplement 7.2). Dreaded risks tend to be global, uncontrollable, involuntary, and hard to reduce. Unknown risks tend to be hard to notice and observe, and to have delayed effects. Both factors featured prominently in the psychometric study of risk perception spearheaded by Slovic.[45]

We then asked coders to read the free-text responses to see whether they would agree with each other and with the authors of the text on whether the risks taken, not taken, or even

regretted were cues of high or low willingness to take risks—that is, whether there was intersubjective agreement on how risk preferences are revealed in experiences and choices. Nine coders read the free-text answers and coded each in triplicate. Coders noted the presence of risk domains, such as investments or health, and more specific risk factors, such as skydiving or divorce. Last, we asked each coder to estimate—based solely on the text responses—the stated risk preference (GRQ) of the respondent.

Results

What risks do people invoke?

Across both studies, 2,510 respondents (72%) gave free-text responses that were sufficiently elaborate to code risk topics (see Supplement 4 for an analysis of nonresponse). The coded topic frequencies for the two free-text questions were highly correlated ($r = 0.94$), so we report summed frequencies in the following (see also Supplementary Tables 7 and 8). Table 1 shows the frequency with which risk domains and subtopics were mentioned.

Table 1. Frequencies with which risk domains and subtopics were mentioned.

Domain	Mentions	Q1	Subtopics
investments	771	418	investment (242), bought home (86), founded company (15), sold home (13)
relationships	760	399	moving (132), conflicts (79), children general (59), speaking out (44), separation (36), pregnant (26), marriage (24), moving in (14), divorce (13), colleagues (10), affairs (7), sticking by (7)
traffic	645	332	car (278), bicycle (172), motorcycle (44), airplane (33), bus (18), train (1)
career	612	321	
safety	437	239	disregarding own frailty (85), working around house and garden (75), going out alone (36), risking being mugged (34), showing moral courage (31), exposure to terrorism (3), fireworks (0), weapons (0)
travel	433	212	
sports	414	233	mountaineering (100), water sports (36), skiing (33), skydiving (23), swimming (19), bungee jumping (8), jogging (7), motor sports (1), shooting sports (0)
health	371	136	surgery (116), drinking (15), immediate health risks: other (14), long-term health risks: other (9), drugs: other (8), sex (7), smoking (7), unhealthy food (7), medication side effects (2), vaccines (1), cannabis (0), GMO food (0), toxins: other (0), pesticides (0), air pollution (0), coffee (0), vaccine avoidance (0)
other	229	144	
gambling	119	59	
crime	37	15	commit misdemeanors (18), commit crime (4)
cataclysm	14	10	terror attack (3), earthquake (1), flooding (0), nuclear waste/war/accidents/fallout (0)

Note. All numbers reflect the number of times a topic was coded from the texts written by our respondents in response to both of the free-text questions. The column Q1 shows the number of mentions in response to the first free-text question (on which risks people thought about).

The topics respondents mentioned frequently tended to be low on the dimensions Unknown and Dread (Figure 2; in addition to the coded categories, we present unigram and bigram word clouds for all responses in Supplement 6.4).

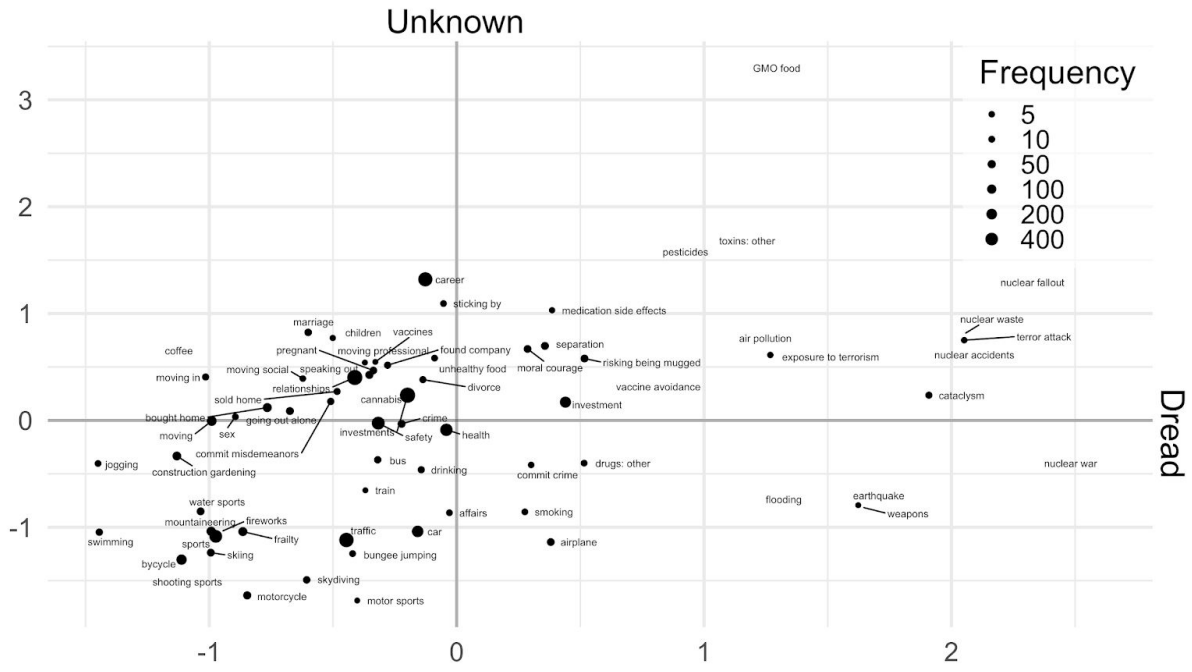

Figure 2. Risk topics are placed in a coordinate system of the Dread (left to right) and Unknown (bottom to top) factors extracted from the risk perception ratings of our online sample. The size and transparency of the dots reflects how often these topics were coded from the responses to the two free-text questions.

When thinking about their risk preferences, respondents focused on common, known risks. We can further characterize the frequently mentioned risks in terms of the rated subdimensions (see supplementary Figure 18): People tended to reference risks that they took *voluntarily* (sports, not terror attacks), that had *observable* consequences (getting on a ladder, not side effects from medication), and which they could *control and prevent* (cars and bikes, not planes and buses).

In line with that pattern, respondents focused on health risks such as surgery and other interventions with immediate consequences, not on chronic risks that have cumulative and delayed effects such as drinking and smoking. The exceptions to these trends were often non-mortality risks: Investment, career, and relationship risks do not always have immediate, knowable consequences. In fact, career and education decisions were the only frequent risk that ranked high on the Unknown dimension. Nobody mentioned the three most unknown topics (according to our online raters): GMO food, pesticides, and “toxins: other.” Respondents almost

never mentioned topics that were dreadful, such as nuclear war or similar cataclysmic events. The most common dreadful topic—terror attacks—was mentioned by only nine respondents.

Which reference frames do people use?

Respondents reported diverse social and temporal reference frames in our two closed-form questions. In both studies, most respondents said they thought of their own experiences and behaviour, or the consequences of their actions, whereas a substantial minority also mentioned comparison with others or what others say (Figure 3). We varied the available response options across the two samples (see Supplement 4). The BASE-II respondents answered an additional question about temporal reference frames; almost all said they thought about the present (78%, n=1,209) or the past (70%, n=1,081), and most of these respondents (52%, n=807) thought about past and present (Figure 4). A substantial fraction of respondents (39%, n=607) also referred to the future, but rarely without thinking about either the past or the present as well (1%, n=20). Some (10%, n=161) respondents additionally endorsed an aspirational reference frame—they thought about how they would like to be—or said they did not think about themselves, but they usually endorsed the more common temporal reference frames too.

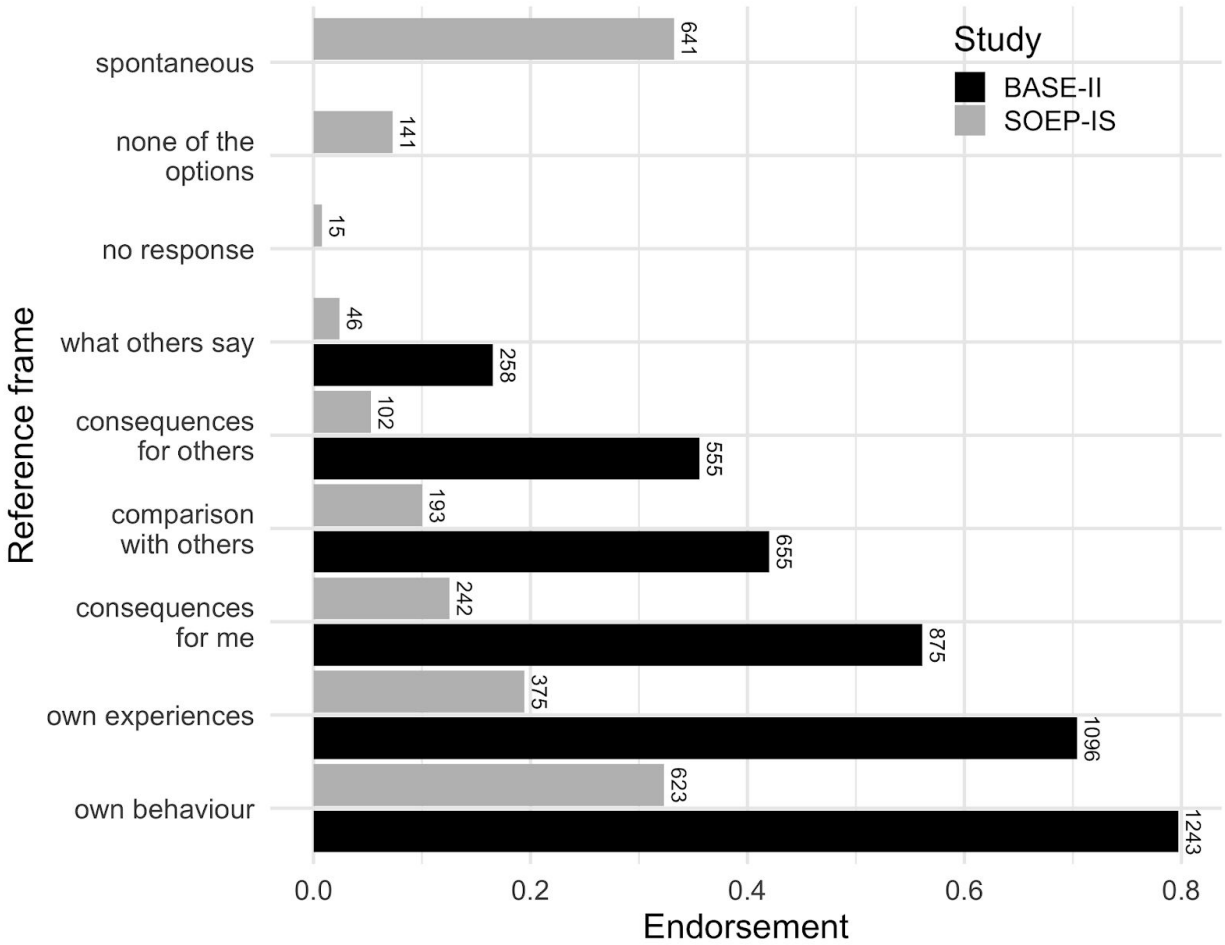

Figure 3. Social reference frames. Compared to SOEP-IS respondents, BASE-II respondents endorsed more options and did not have the option to say they responded spontaneously, or based on something else. The options that were common to both studies were similar in rank.

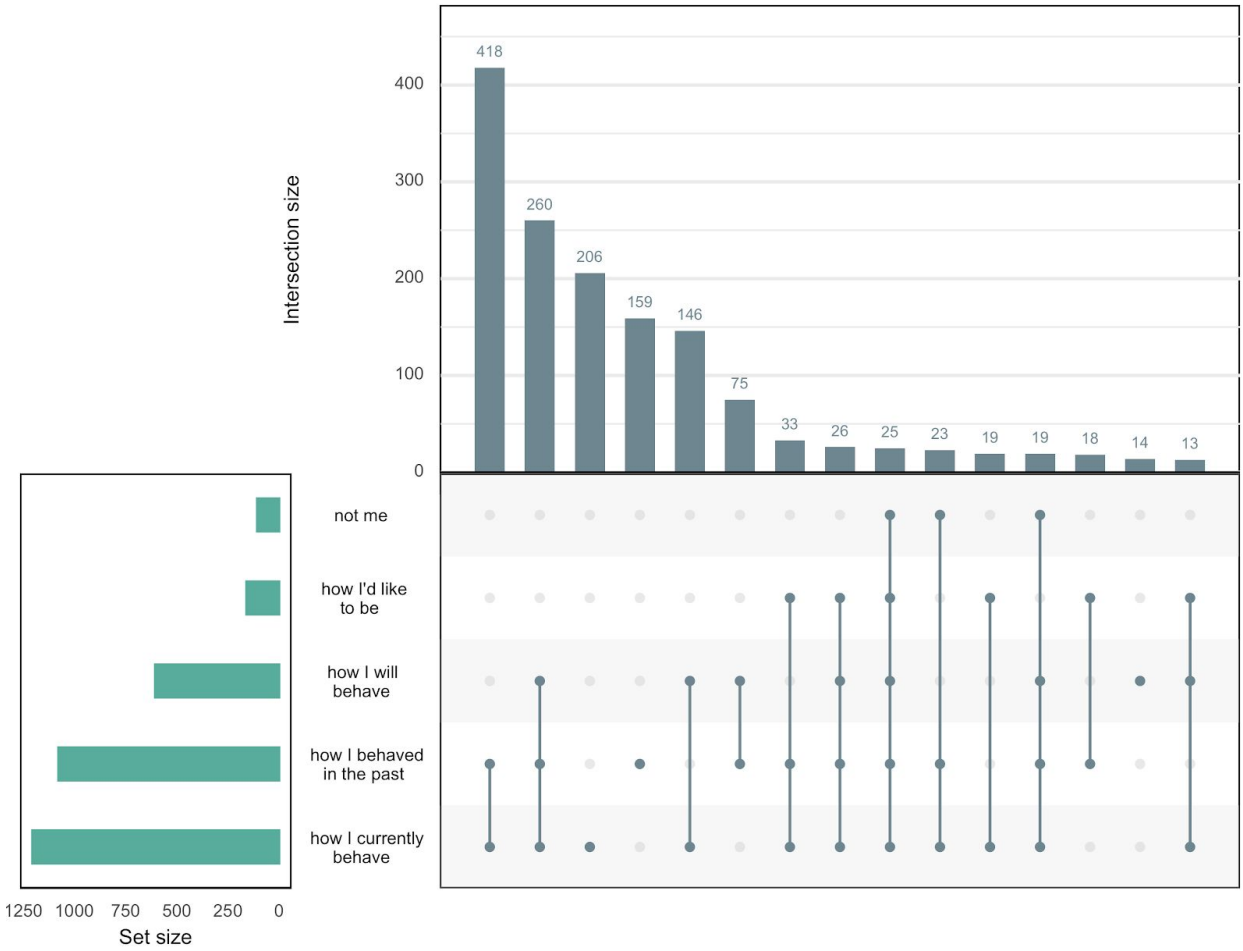

Figure 4. Temporal reference frames. This UpSet plot [47] shows the frequency of endorsing one or several options in the question about temporal reference frames in the BASE-II study. The lower left panel shows simple counts, whereas the top panel shows how options were combined. Only the fifteen most common combinations are shown here.

Do people think about risks they took, avoided, or both?

Among those who mentioned codeable risks, most (53%, n=1,129) respondents clearly mentioned risks they took, and only 2% mentioned risks they avoided. For the remainder of responses, it was unclear whether risks were taken or avoided (32%), no two coders agreed (12%), or respondents wrote about risks that others took (1%). Crime, gambling, and investment risks were mentioned as avoided risks more frequently than the average risk (9%, 3%, and 3%, respectively).

BASE-II respondents were asked whether the risks they had taken in the last year had been worthwhile. Of those respondents who listed a risk taken in the last year, most reported that it was worth taking the risks (68%, n=709) or that the risks were partially worthwhile (11%). A total of 3% gave different answers for different risks, and 4% said it was too soon to tell whether it had been worth taking the risk. Only 9% said clearly that taking the risk had not been worth it, and 1% said they did not know. For 4% of responses no two coders agreed. Compared to the average level of regret, respondents appeared to particularly regret risks taken in the domains of gambling (26% of cases when gambling was the topic), crime (17%), and traffic (14%), whereas few regretted taking risks related to relationships (5%), sports (4%), their career or educational decisions (3%), and travel (1%).

Do age and gender affect the risks people refer to?

On average, men were more likely to mention risks of injury such as traffic and sports risks. Women mentioned relationship and travel risks more often than men. Older people—women and men alike—rarely mentioned career and educational decisions or sports, but increasingly mentioned traffic, health, and safety risks (Figure 5). Young men were most likely to mention gambling; otherwise age trends were largely parallel for men and women. Age and gender differences in reference frames were less pronounced than topic differences (see Supplementary Figures 8–11).

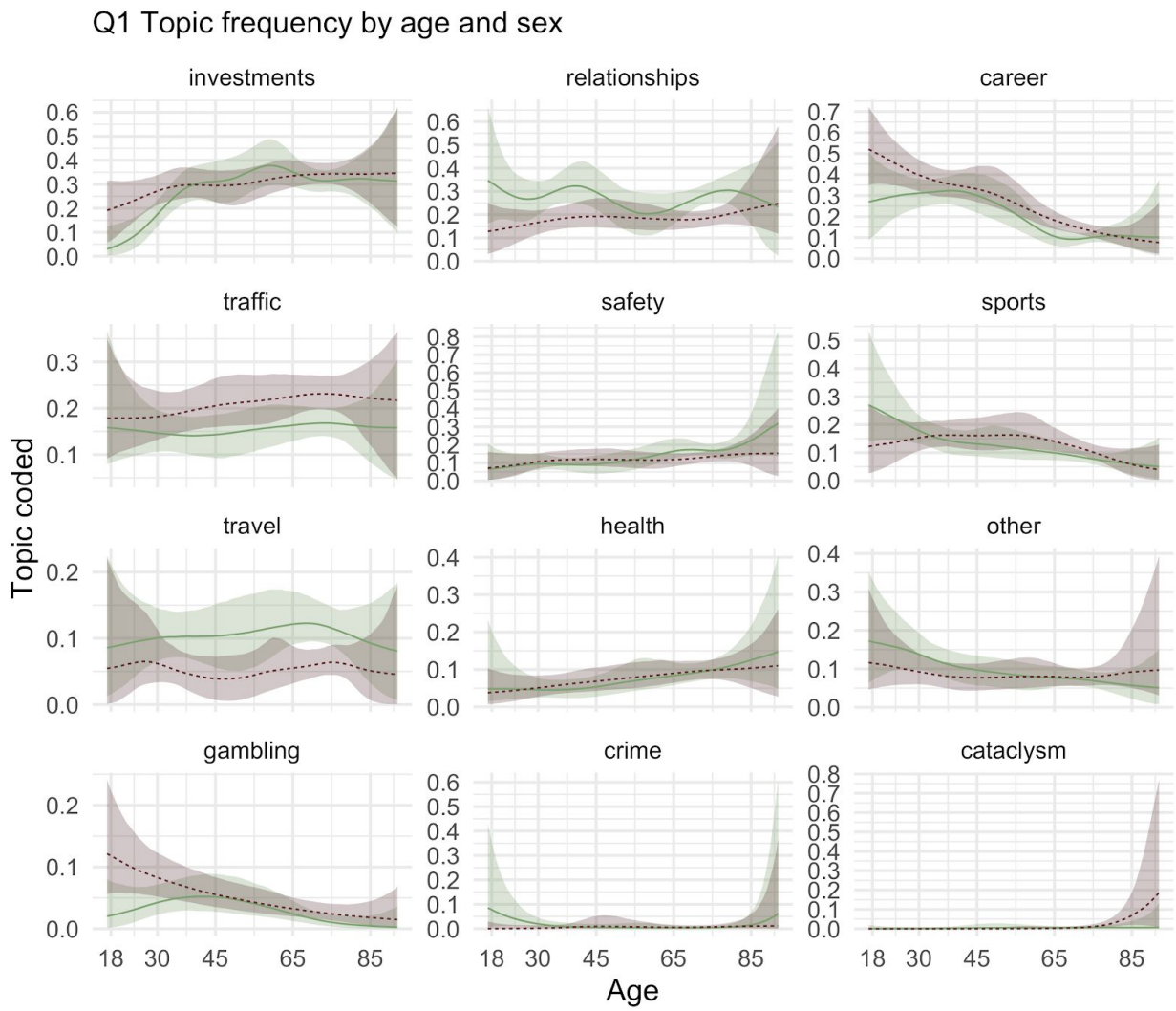

Figure 5. Age trends and gender differences in risk domains coded based on what people thought about when answering the General Risk Question. The lines show regression splines by gender with shaded 95% credible intervals. Dashed lines indicate men, solid lines women. The BASE-II and SOEP-IS samples were pooled and a contrast-coded dummy for study was adjusted for.

Can independent third parties agree on what people's experiences say about their preferences?

We found that coders could—based solely on the texts—estimate the stated risk preference (on a scale from 0 to 10) of the text's author above chance level by using cues such as the number

of risks, whether risks were seen as worthwhile, or whether risks were avoided (see Supplementary Table 23). The zero-order correlation between stated preferences and mean coder estimates was 0.27 (95% CI [0.23;0.31]) and was well-described by a linear function (see Figure 6). Coders agreed not only with the respondents, but also with one another: When weighted by the coders' confidence, the intraclass correlation (ICC) was .63 (unweighted ICC .43), showing substantial agreement across coders. When coders were more confident, their judgments were also more accurate (see Supplementary Figure 22). Coders only minimally underestimated respondents' risk preferences on average and less so when coders were confident (by 0.14 points, see Supplement 8.2). Coders tended towards the mean, overestimating low preferences for risk and underestimating high preferences. This tendency was more pronounced when coders were less confident in their judgment.

We carried out a social judgment analysis[48,49] to determine which cues coders used to infer stated risk preferences and how validly these cues could predict respondents' stated preferences. Results showed that coders generally used valid cues (i.e., cues such as the number of risks which predicted both coder judgments and respondents' stated preferences; r=.74 between predicted judgments and predicted outcomes). However, coders also used some invalid cues. For instance, coders rated those who responded vaguely as lower in risk preference, even though vagueness was not predictive of stated risk preference (see Supplement 8.7). A pastiche, to preserve anonymity, of a text that received the lowest rating would be: “I always keep my head out of things, and only take credits with fixed interest rates. In the last year, I tried a new restaurant.” A pastiche for someone who received the highest rating would be “I thought about races on the motorway, and cheating on my partner. In the last year, I travelled abroad without having any money.”

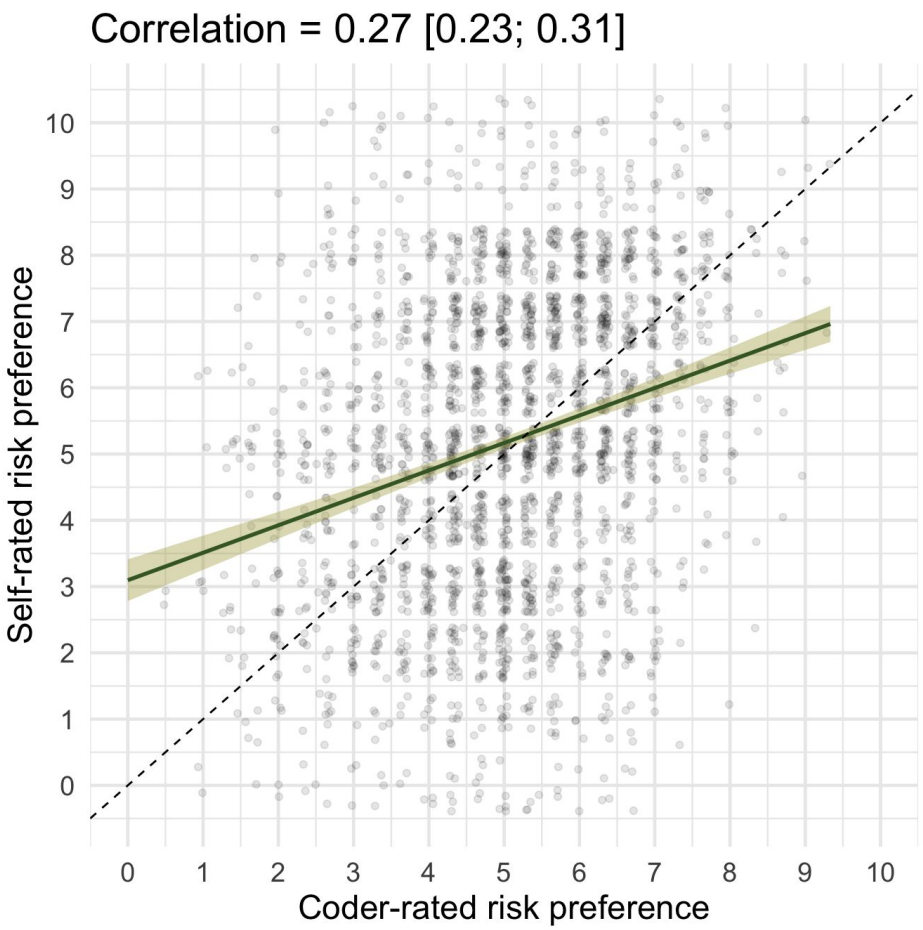

Figure 6. The green line shows a linear regression fit with the 95% confidence interval shaded. Along the dashed line, coder and self-ratings matched. Points were jittered slightly to reduce overplotting.

We also tested whether the coders could infer risk preferences from the texts equally well for respondents with different ages and genders to see whether idiosyncrasies in risk perception across age groups and gender might decrease the validity of stated preferences. We jointly tested several potential modulators of coders' ability to infer risk preferences—study, respondent's age, respondent's gender, and the coder having the same gender as the respondent—to separate their contributions to accuracy while adjusting for the number of characters written. This model was necessary due to variations between the two studies; for example, BASE-II respondents wrote more characters and were older on average than SOEP-IS respondents. In this model, accuracy did not differ depending on the respondents' age, gender, or the coder's gender being the same as the respondent's. However, BASE-II respondents were rated more accurately (i.e., coders' evaluations matched respondents'

self-evaluations) by coders (*r* = .33 vs. *r* = .21 in SOEP-IS; see also Table 2.), fitting the finding that considering risks worthwhile (this question was not asked in SOEP-IS) was a valid cue in the social judgment analysis.

Table 2. Results from a distributional regression.

Predictor	Estimates	CI (95%)
Intercept	4.27	3.66; 4.89
Stated risk preference	0.15	0.13; 0.18
σ – Intercept	0.23	-0.07; 0.51
σ – BASE-II participant	-0.08	-0.13; -0.03
σ – Male gender	-0.01	-0.05; 0.03
σ – Coder has same gender	-0.01	-0.06; 0.03
σ – Age (in decades)	0.00	-0.01; 0.02
σ – log10 (nr. of characters)	0.05	0.03; 0.08
sd(Respondent-Intercept)	1.06	1.02; 1.11
sd(Coder-Intercept)	0.80	0.46; 1.45
sd(σ - Intercept)	0.42	0.24; 0.76

Note. The model was fit in brms.[50] We let respondents' stated risk preferences predict the coder ratings of risk preference and let several moderators jointly predict the error term (σ) in order to disentangle their contributions. BASE-II participants were rated more accurately, when adjusting for the effects of age, gender, coder gender, and number of written characters. The model includes 2,293 respondents rated 6,863 times by nine coders (~3 each).

Discussion

To investigate how stated preferences can be valid, we asked respondents to explain their answers to a general question about their risk preferences.[6] Our results show that people establish a shared reference frame by seeing what preferences are revealed in the risks they themselves took, avoided, and regretted. We argue that this self-judgment taps into an ability for social judgment[29,31] that people need to quickly assess whether someone will be a steadfast ally or an unpredictable enemy.[32] With just a brief glimpse into our respondents' self-perceptions, our coders were able to infer their stated risk preferences to a significant extent. Coders did even better when, as in the BASE-II study, they had access to information about respondents' experiences of regret. We argue that self-judgments of risk preferences take into account not just actions, but also situational constraints and internal states such as experiences of regret, or need.

The risks people thought about were highly heterogeneous, but most respondents focused on voluntary behaviours and decisions with risk of easily observable harm, including physical, financial, and social risk. Major life decisions, especially risks taken in relationships, investments, and careers were often mentioned. Cumulative and delayed risks of harm, such as smoking or unprotected sex, were mentioned only infrequently. Passively tolerated sources of risk from technology or natural hazards were almost never mentioned. It seems that when people consider which actions reveal their risk preferences, they think of more diverse actions than the ones experimental economists and psychologists use in the laboratory. Gambling, the most common laboratory measure of risk preferences, was mentioned only rarely, and unlike more commonly mentioned risks it was avoided and regretted more often. Seen through the eyes of our respondents, gambling is an odd risk: The precisely defined risk, the possibility of avoiding gambling entirely, and the frequency of regret all make it different from the more commonly mentioned risks taken in relationships, investments, health, and careers. In contrast, the widely used DOSPERT questionnaire[51] asks about a list of hypothetical behaviours that may better capture the diversity of risks people have in mind, in terms of both risk domains and size of stakes. It includes everyday behaviours such as not wearing a seatbelt, more rare behaviours like having an affair, and rare but important events like choosing a more enjoyable but less secure career. In our data, relationship and career risks were also prominent, especially among

the biggest risks faced in the previous year (see also Supplementary Note 1). These risk domains are amongst those highest on the Unknown dimension of the Slovic system: Decisions about whether to marry, divorce, move, quit a job, or study a particular subject are highly uncertain and can seriously alter a life's trajectory. Respondents realized this and frequently mentioned such decisions with very large stakes—which may indeed reveal more about their own risk preferences than risks with low stakes as are typical in the laboratory. It is possible that preferences were not only revealed through these decisions but also shaped by their consequences: As people learn through trial and error, their preferences mature over the lifespan.[52]

The difficulty of building revealed risk preference tasks in domains like relationships makes representative designs, which capture the ecology of risks, less likely in the laboratory.[4,48] Much research operates under the assumption that the extrapolation from small to big risks is possible[4,13]—that the person who gambles in a laboratory lottery will also gamble with their life and happiness. However, this assumption may not hold. More work needs to be done to account for the older evidence showing that people are more risk averse when facing high financial stakes,[53] mounting evidence of the low criterion validity of revealed risk preference tasks,[5,54] and recent work finding that hypothetical lotteries are workable proxies of incentivised ones.[55] Potentially, any shared validity between hypothetical lotteries and stated preferences results from the same process: People look to their past actions and experiences to construct a response to an abstract decision.[22,24,56] This process may also explain the validity of the DOSPERT questionnaire, in which all behaviours are hypothetical and people predict their own behaviour. Still, even the 30–40 items of the DOSPERT questionnaire cannot capture all the idiosyncratic, yet pertinent risks our respondents listed (e.g., “buying a horse and never telling your partner”), but people could draw on idiosyncratic experiences to reasonably predict their own behaviour in standardised hypothetical situations. Perhaps the DOSPERT questionnaire additionally works as an aid for dialectical bootstrapping,[57] helping people to come up with several responses that reflect their true preference plus noise, which can then be averaged for increased reliability (see also Supplementary Note 2).

Because our coders could, to a significant extent, infer respondents' risk preferences from the texts, we know the texts contained valid cues, such as the number of risks and whether risks were avoided or regretted. In fact, the correspondence between coder ratings and stated

preferences in our study (r = .27) was similar to the correspondence between risk perceptions in self-ratings and ratings by close informants (*r*s = .25-.46[58]) and to the correspondence for decisions between lotteries (*r* = .31) between two household members.[59] It was also close to the agreement between self and other ratings among Facebook friends for personality traits.[60] Despite being very brief—a median of 10 words—the texts contained condensed, pertinent information. Our social judgment analysis showed that coders relied on cues such as regret, the number of risks listed for the last 12 months, and risk avoidance. They also took note of specific risky activities, such as motorcycling and sports, and correctly inferred that respondents who listed investments as a risky activity had stated lower risk preferences.

The topics respondents thought about differed by age and gender. For example, an elderly respondent listed “getting into the bathtub” as a risk, which might not strike younger respondents as particularly threatening. More generally, older respondents were more likely to mention risks in health and traffic, and less likely to focus on their career or gambling. Gender and age differences in risk perception and conception (i.e., focusing on favourable or unfavourable outcomes[61]) might provoke doubt that there is a common denominator that allows for comparing stated risk preferences across age groups and genders. We think it is actually the other way around: Risk perception and conception are cues to people's risk preference too.[61,62] In initial support of this notion, our coders—aged between 23 and 36—were equally accurate when inferring the preference of older respondents or those of the opposite gender. Given that people can agree on perceptions of risk,[45,62] as we found in our online sample, they can also agree on what taking specific risks implies for somebody's risk preferences. This interpretation leads to a more optimistic conclusion than the widespread idea that people always anchor themselves to a social reference group (which would change according to age, location, and time). Indeed, only a minority of our respondents said they used social comparison; most said they simply thought about their experiences and behaviours. This result may explain why, in apparent conflict with a cognitive model of personality judgments,[63] specifying reference groups reduced predictive validity in a study of conscientiousness.[41] If most people do not naturally tend to compare themselves to a reference group, they may do worse when required to do so. Much of the literature has focused on finding out whether questions could be improved, by specifying their frame of reference,[41,42] reference groups,[63,64] examples,[40] or specific behaviours,[65,66] or by generally reducing temporary, fluctuating influences.[27,28] In risk preference research, Blais and Weber[51] attempted to remove any part played by differences in risk perception.

Counterintuitively, leaving self-report questions fairly broad and vague may sometimes improve validity, as long as people understand the question and can draw on relevant experiences. A comprehensive single item may allow people to use their ability of social perception to draw on their most pertinent and diagnostic information, thereby outweighing improvements from further constraints on the question that would make answering it less like social perception.

We conclude that what many researchers feel is a weakness of stated preferences might actually be a strength.[15] The fairly vague, almost projective nature of a comprehensive single item allows people to refer back to their diagnostic memories and behaviours using a well-honed human capacity for social perception. People with different risk perceptions and conceptions could be problematic for the intersubjective comparability of their answers,[61] but we find that people (our coders) can agree on what risky behaviours imply for someone's risk preference, irrespective of age and gender. The shared social perception of risks fosters agreement, comparability, and validity of risk preferences too. This does not mean that self-reports can be trusted to screen for risk-seekers when trying to recruit responsible financial managers. People could still lie—just as they could in typical laboratory tasks, where stakes are generally low.

Far from “cheap talk,” self- and informant-reports are based on informative and diagnostic cues and permit people to apply the full might of social perception to themselves. These results suggest that researchers in economics and psychology can learn from the experts on person perception: their study participants. By inferring their risk preferences from their diagnostic behaviours and experiences, people essentially adopt the logic of the revealed preference approach—namely, that otherwise unobservable preferences reveal themselves in behaviour. Ironically, economics’ revealed preference approach, which has recently encountered many difficulties in the experimental literature studying behavioural risk measures, appears to have found new significance in research on stated risk preference—the method that the revealed preference approach has deemed of dubious value.

Materials and Methods

All questions and materials needed to reproduce the study have been shared on Open Science Framework (OSF) at osf.io/eun4r/. The main questions can be found in Supplementary Note 3.

The stated preferences were collected in the 2017 interim wave of the Berlin Aging Study II (BASE-II[43]) and the 2017/2018 wave of the SOEP innovation sample (SOEP-IS[44]). Both studies are age-heterogeneous longitudinal panel studies. SOEP-IS aims to representatively sample private households in Germany, whereas BASE-II is a convenience sample of younger and older adults from Berlin, Germany. Participants in both studies had already answered the general and domain-specific risk questions in previous waves. In the 2017/2018 wave, 3,493 respondents answered the GRQ and 3,089 answered several questions that elicited free-text source reports. Both studies have been documented on https://paneldata.org. Fieldwork for SOEP-IS started in September 2017 and ended in February 2018. Questionnaires for BASE-II were mailed out at the beginning of November 2017; data collection ended in January 2018. The online rater sample was recruited from the psytests.de and the psyweb.uni-muenster.de online panels from April to August 2018. Participants could win one of 50 Amazon coupons worth €25 each in a lottery. The coders were recruited from the participant pool of the Max Planck Institute for Human Development and were paid €180 each. Descriptive statistics for all samples are summarised in Table 3. The anonymised data for the online rating study is available on OSF. The SOEP-IS data can be obtained from the SOEP re-analysis archive; the BASE-II data can be obtained from the BASE-II Steering Committee.

Table 3. Demographic statistics for the three samples.

	SOEP-IS (n=1,928)		**BASE-II** (n=1,569)		**Online Raters** (n=944)		**Coders** (n=9)
	Mean (SD)	Missing	Mean (SD)	Missing	Mean (SD)	Missing	Mean (SD)
Age	53.4 (18.6)	0	66.6 (15.9)	0	46.8 (17.6)	272	27.9 (4.4)
Male	47%	0	48%	0	39%	281	56%
Gen. Risk Q.	4.6 (2.4)	0	5.2 (2.3)	4	4.4 (2.1)	123	
No. of words	7.5 (8.0)	274	18.0 (15.5)	138			
Text length	51 (51)	274	135 (106)	134			
Codeable topics Q1	46%	0	80%	0			
Codeable topics Q2	40%	0	67%	0			

Note. SD=Standard deviation. There were no missing values for the coders. A subsample of n=825 online raters rated the risks (n=119 ended the study before the ratings).

Measures

Stated preferences

Stated preferences were measured using the General Risk Question (GRQ[6]). After the respondents answered this question, they were asked a series of follow-up questions. We slightly reduced the number of questions for SOEP-IS compared to BASE-II to fit the time requirements of the panel. In both studies, the first follow-up question was “Which events, behaviour, or persons did you think about when you indicated a number for your risk preference?” Participants could check multiple options: “own experiences,” “own behaviour,” “my behaviour compared to others,” “the consequences of my behaviour for me,” “the consequences of my behaviour for others,” and “thought about what people around me say about my risk preference.” In SOEP-IS, respondents could additionally choose from several nonresponse options: “gave my answer spontaneously without deliberating a great deal,” “none of these,” and “no answer.” In BASE-II, a second multiple choice question asked respondents whether they thought about one or more of the following options: “how I presently behave in my

day-to-day life," "how I behaved in the past," "how I will behave in the future," "how prepared for risks I would like to be," and "did not think about myself." In both studies, the closed-form questions were followed by free-text questions. They were "Which concrete experiences or behaviours—it does not matter if yours or others'—did you think about? Please give keywords" and "In which situations in the last 12 months were you prepared to take risks? List up to three situations in which you took the biggest risks. Keywords suffice." In BASE-II only, respondents were then asked, "And were the risks worth it?"

The BASE-II respondents filled out paper-and-pencil questionnaires and returned them by mail. They were given four lines to write on for each free-text question. Their responses were later transcribed by student assistants at the Humboldt University Berlin. In SOEP-IS, respondents answered verbally and the interviewer transcribed their answers during computer-assisted personal interviewing. BASE-II respondents gave valid and elaborate answers to the free-text questions more frequently than did the SOEP-IS participants: 92%, compared to 86% (*n*s=1,435; 1,654) answered at least one of two free-text questions. BASE-II respondents wrote a median number of 106 characters; the median for SOEP-IS respondents was 35 characters. Texts by BASE-II respondents were sufficiently informative to code risk topics for 1,248 responses to the question asking them to explain their thinking for the stated preferences, and for 1,056 responses to the question asking about risks taken in the last year. Given the shorter responses in SOEP-IS, topics were codeable only for *n*s=890/773 free-text responses (see also Supplementary Section 4).

Text coding

The texts written by the BASE-II and SOEP-IS participants were hand-coded by a set of nine coders (aged 23–36, four women) over multiple days. We randomly divided the full-text answers into two sets of 1,000 and one set of 1,059 answers. The coding scheme was derived through a mixture of a deductive approach (risk factors listed in the literature[45]) and an inductive approach (further risk factors mentioned in the texts). For initial training, all coders coded a set of the same 50 texts. Afterwards, the coding scheme was refined and agreement was checked according to Fleiss' Kappa. Points of disagreement about the scheme between coders were resolved by the first author (RCA). For the remainder of the texts, three coders coded each text. Coders tended to agree on the presence of risk domains; Fleiss' Kappas were above .70 for all

coder groups (see Supplementary Table 20) and all risks except safety and crime (κ ≥ .49, because coders could not always agree whether respondents were perpetrators or victims of crime), and cataclysms (κ = .00–.61, but this category was very rare). They also noted whether the texts mentioned risks that were taken or avoided (here, agreement was only slight: κ = .04–.18) as well as whether respondents thought the risk had been worthwhile (κ = 0.71–0.77).

Coders saw all the answers to the free-text questions given by a respondent simultaneously in case the answers referenced each other. They did not see the answers to the closed-form questions or other identifying characteristics. First, coders judged whether meaningful topics or situations were mentioned in the response. If not, they could code whether the response was gibberish, a statement of absence, or similar. The coders then coded the presence of the topics from the coding scheme such as health and relationships for each of the two free-text questions. Some topics included subcategories (e.g., operations or divorce) that could be coded (see Supplementary Tables 25-26). For the first question, which asked respondents to explain their thinking for their stated preferences, coders noted whether the situations and events described had a risk prevention or promotion focus (the second question was explicitly about risks taken in the last year and therefore could not be codified according to risk prevention or promotion). For the question asking whether risks were worthwhile, which appeared only in BASE-II, coders noted whether the respondents thought the risk had been worthwhile or whether they were unable to tell so far (e.g., long-term financial risks). Finally, the coders rated the respondent on the same GRQ that the respondent had answered. For our analyses, we chose the consensus value given by the coders (i.e., the coding made by at least two coders) or the mean for continuous values. For the 50 texts that we used to train coders, we omitted the data from the first six coders before aggregation to keep the procedure comparable for all texts.

Analyses

Our data processing code, statistical analyses, and detailed results are reproducibly documented on OSF (osf.io/eun4r/).

Online rating of risk perceptions

Online participants rated the risk factors or behaviours from our coding scheme (e.g., moving in together and smoking) on 22 dimensions (e.g., observability and reducibility). To find out how

reliable the average ratings were, we computed average ICCs for each dimension for an average of 17 aggregated ratings, which was the lowest number of ratings any risk topic had received (median = 37). Average ICCs ranged from .73 (whether risks were known to science) to .97 (whether these were risks for social position). These ICCs are lower bounds, as most risks were rated by more than 17 raters. Because it is not possible to meaningfully answer questions such as “Are health risks known to science?” the online sample did not rate broad and vague risk domains such as health and traffic; instead, we averaged the ratings of the constituent subtopics to arrive at values for the risk domains. To lay out the risk factors on a familiar map for our readers, we extracted the factors Dread and Unknown according to a confirmatory specification based on 16 dimensions from Slovic[45]. We could approximately replicate the coordinate system positions of risks in Slovic[45], fulfilling our limited aim, but—probably because we had added non-mortality, social risks—fit indices fell short (see Supplement 7). Owing to a programming error, the topics “gambling”, "travel", and "surgery" were not rated by the online sample and are therefore missing in Figure 2.

Coder-estimated risk preferences

Coders had indicated whether the text contained direct hints to the authors’ gender, age, or place of residence, such as, “My husband lost at bingo in our retirement home in Munich.” Because such hints might serve as cues to the stated risk preference, given age and sex differences in risk preferences, but would be unrelated to risk conceptions per se, we restricted the main analysis to the majority (97%, n = 2,310) of texts which contained no direct hints. Even indirect hints, such as considering “getting into the bathtub” a risk, seemed to play little role: accuracy was not attenuated when we adjusted for respondent age and gender (see Supplement 8.1).

Coders could tell when they had usable information. Accuracy was r = .06 when coders said they were guessing, but r = .45 when they had maximal confidence (see Supplement 8.5). Coders did not learn to judge more accurately with practice, which we expected given that they received no feedback.

Acknowledgements

The authors thank Jürgen Schupp for helping to design the survey questions we applied, Larissa Samaan for translating the word clouds, Jann Wäscher for organizing the coder appointments, the members of the Center for Adaptive Rationality for their helpful comments on earlier versions of this work, our coders, and all participants of BASE-II, SOEP-IS, and our online study. The BASE-II research project is supported by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) under grant numbers #16SV5536K, #16SV5537, #16SV5538, #16SV5837, #01UW0808, #01GL1716A, and #01GL1716B. Additional contributions were made from GGW's Max Planck Fellowship Grant. The SOEP-IS survey is funded via the Leibniz Gemeinschaft by the Federal Government of Germany and the Federal State Berlin. We thank Deb Ain for scientific editing. All remaining errors are ours.

Author contributions

GGW and RH together with TD designed the questions for the SOEP-IS and BASE-II studies. RCA designed and executed the coding study and the online rating study. JD prepared the BASE-II data. MB conducted the text mining and generated the related figures. RCA analyzed all data. RCA, RH, and GGW wrote the first manuscript draft. All authors critically and substantively revised the manuscript.

References

1. Mata, R., Frey, R., Richter, D., Schupp, J. & Hertwig, R. Risk Preference: A View from Psychology. *J. Econ. Perspect.* **32**, 155–172 (2018).
2. van Oers, K., Drent, P. J., de Goede, P. & van Noordwijk, A. J. Realized heritability and repeatability of risk-taking behaviour in relation to avian personalities. *Proc. Biol. Sci.* **271**, 65–73 (2004).

3. Steinberg, L. *et al.* Age differences in sensation seeking and impulsivity as indexed by behavior and self-report: evidence for a dual systems model. *Dev. Psychol.* **44**, 1764–1778 (2008).
4. Hertwig, R., Wulff, D. U. & Mata, R. Three gaps and what they may mean for risk preference. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* (2019) doi:10.1098/rstb.2018.0140.
5. Frey, R., Pedroni, A., Mata, R., Rieskamp, J. & Hertwig, R. Risk preference shares the psychometric structure of major psychological traits. *Science advances* **3**, e1701381 (2017).
6. Dohmen, T. *et al.* Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences. *J. Eur. Econ. Assoc.* **9**, 522–550 (2011).
7. Caliendo, M., Fossen, F. & Kritikos, A. S. Personality characteristics and the decisions to become and stay self-employed. *Small Bus. Econ.* **42**, 787–814 (2014).
8. Caliendo, M., Fossen, F. M. & Kritikos, A. S. Risk attitudes of nascent entrepreneurs–new evidence from an experimentally validated survey. *Small Bus. Econ.* **32**, 153–167 (2009).
9. Falk, A. *et al.* Global Evidence on Economic Preferences. *Q. J. Econ.* **133**, 1645–1692 (2018).
10. Friedman, M. & Savage, L. J. The Utility Analysis of Choices Involving Risk. *J. Polit. Econ.* **56**, 279–304 (1948).
11. Friedman, D., Isaac, R. M., James, D. & Sunder, S. *Risky curves: On the empirical failure of expected utility*. (Routledge, 2014).
12. Harrison, G. W. & Rutström, E. E. Chapter 81 Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods. in *Handbook of Experimental Economics Results* (eds. Plott, C. R. & Smith, V. L.) vol. 1 752–767 (Elsevier, 2008).
13. Charness, G., Gneezy, U. & Imas, A. Experimental methods: Eliciting risk preferences. *J.*

Econ. Behav. Organ. **87**, 43–51 (2013).

14. Holt, C. & Laury, S. Risk Aversion and Incentive Effects. (2002) doi:10.2139/ssrn.893797.
15. Dana, J., Atanasov, P., Tetlock, P. & Mellers, B. Are markets more accurate than polls? The surprising informational value of 'just asking'. *Judgm. Decis. Mak.* **14**, (2019).
16. Tynan, M. The Domain-Specific Risk-Taking Scale lacks convergence with alternative risk-taking propensity measures. (Iowa State University, 2018). doi:10.31274/etd-180810-6107.
17. Harden, K. P. *et al.* Beyond dual systems: A genetically-informed, latent factor model of behavioral and self-report measures related to adolescent risk-taking. *Dev. Cogn. Neurosci.* **25**, 221–234 (2017).
18. Charness, G., Garcia, T., Offerman, T. & Villeval, M. Do measures of risk attitude in the laboratory predict behavior under risk in and outside of the laboratory? (2019) doi:halshs-02146618v2.
19. Pedroni, A. *et al.* The risk elicitation puzzle. *Nature Human Behaviour* (2017) doi:10.1038/s41562-017-0219-x.
20. Pachur, T., Mata, R. & Hertwig, R. Who dares, who errs? Disentangling cognitive and motivational roots of age differences in decisions under risk. *Psychol. Sci.* **28**, 504–518 (2017).
21. Vieider, F. M. *et al.* Common Components of Risk and Uncertainty Attitudes Across Contexts and Domains: Evidence from 30 Countries. *J. Eur. Econ. Assoc.* **13**, 421–452 (2015).
22. Lichtenstein, S. & Slovic, P. *The Construction of Preference*. (Cambridge University Press, 2006).
23. Jarecki, J. B. & Wilke, A. Into the black box: Tracing information about risks related to 10

evolutionary problems. *Evolutionary Behavioral Sciences* (2018).

24. Steiner, M., Seitz, F. & Frey, R. Mapping the Cognitive Processes Underlying Self-Reported Risk-Taking Propensity. *PsyArXiv* (2019).
25. Schwarz, N. Self-reports: How the questions shape the answers. *Am. Psychol.* (1999).
26. Sedikides, C. Assessment, enhancement, and verification determinants of the self-evaluation process. *J. Pers. Soc. Psychol.* **65**, 317–338 (1993).
27. Schimmack, U. & Oishi, S. The influence of chronically and temporarily accessible information on life satisfaction judgments. *J. Pers. Soc. Psychol.* **89**, 395–406 (2005).
28. Schimmack, U., Diener, E. & Oishi, S. Life-satisfaction is a momentary judgment and a stable personality characteristic: the use of chronically accessible and stable sources. *J. Pers.* **70**, 345–384 (2002).
29. Vazire, S. Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *J. Pers. Soc. Psychol.* **98**, 281–300 (2010).
30. Sun, J. & Vazire, S. Do people know what they're like in the moment? *Psychol. Sci.* **30**, 405–414 (2019).
31. Bem, D. J. Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychol. Rev.* **74**, 183–200 (1967).
32. Fessler, D. M. T., Tiokhin, L. B., Holbrook, C., Gervais, M. M. & Snyder, J. K. Foundations of the Crazy Bastard Hypothesis: Nonviolent physical risk-taking enhances conceptualized formidability. *Evol. Hum. Behav.* **35**, 26–33 (2014).
33. Bem, D. J. Self-perception theory. in *Advances in experimental social psychology* vol. 6 1–62 (Elsevier, 1972).
34. Barclay, P., Mishra, S. & Sparks, A. M. State-dependent risk-taking. *Proc. Biol. Sci.* **285**, (2018).

35. Mishra, S., Barclay, P. & Sparks, A. The relative state model: integrating need-based and ability-based pathways to risk-taking. *Pers. Soc. Psychol. Rev.* **21**, 176–198 (2016).
36. Watson, N. & Wooden, M. P. The HILDA Survey: a case study in the design and development of a successful Household Panel Survey. *Longit. Life Course Stud.* **3**, 369–381 (2012).
37. University of Essex, Institute for Social and Economic Research. Understanding Society: Waves 1-8, 2009-2017 and Harmonised BHPS: Waves 1-18, 1991-2009. (2018) doi:10.5255/UKDA-SN-6614-12.
38. Wagner, G. G., Göbel, J., Krause, P., Pischner, R. & Sieber, I. Das Sozio-oekonomische Panel (SOEP): Multidisziplinäres Haushaltspanel und Kohortenstudie für Deutschland--Eine Einführung (für neue Datennutzer) mit einem Ausblick (für erfahrene Anwender). *AStA Wirtschafts-und Sozialstatistisches Archiv* **2**, 301–328 (2008).
39. Linnér, R. K. *et al.* Genome-wide study identifies 611 loci associated with risk tolerance and risky behaviors. *bioRxiv* 261081 (2018) doi:10.1101/261081.
40. Tourangeau, R., Sun, H., Conrad, F. G. & Couper, M. P. Examples in open-ended survey questions. *Int J Public Opin Res* **29**, 690–702 (2017).
41. Credé, M., Bashshur, M. & Niehorster, S. Reference group effects in the measurement of personality and attitudes. *J. Pers. Assess.* **92**, 390–399 (2010).
42. Schmit, M. J., Ryan, A. M., Stierwalt, S. L. & Powell, A. B. Frame-of-reference effects on personality scale scores and criterion-related validity. *J. Appl. Psychol.* **80**, 607–620 (1995).
43. Bertram, L. *et al.* Cohort profile: The Berlin Aging Study II (BASE-II). *Int. J. Epidemiol.* **43**, 703–712 (2014).
44. Richter, D. & Schupp, J. SOEP Innovation Sample (SOEP-IS) — Description, structure and documentation. (2012) doi:10.2139/ssrn.2131214.

45. Slovic, P. Perception of risk. *Science* **236**, 280–285 (1987).

46. Carson, R. T., Horowitz, J. K. & Mellissinos, M. *The Relationship between Desire to Reduce Risks and Factor Scores for Environmental Risks*. https://ideas.repec.org/p/ags/umdrwp/197629.html (1989).

47. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).

48. Brunswik, E. Perception and the representative design of experiments. Berkeley. (1956).

49. Cooksey, R. W. Judgment analysis: Theory, methods, and applications. *Judgment analysis: Theory, methods, and applications.* xv, 407–xv, 407 (1996).

50. Bürkner, P.-C. brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* **80**, (2017).

51. Blais, A.-R. & Weber, E. U. A domain-specific risk-taking (DOSPERT) scale for adult populations. *Journal of Judgment and Decision Making* **1**, 33–47 (2006).

52. Josef, A. K. *et al.* Stability and change in risk-taking propensity across the adult life span. *J. Pers. Soc. Psychol.* **111**, 430–450 (2016).

53. Binswanger, H. P. Attitudes Toward Risk: Experimental Measurement in Rural India. *Am. J. Agric. Econ.* **62**, 395–407 (1980).

54. Galizzi, M. M., Machado, S. R. & Miniaci, R. Temporal Stability, Cross-Validity, and External Validity of Risk Preferences Measures: Experimental Evidence from a UK Representative Sample. *Cross-Validity, and External* (2016) doi:10.2139/ssrn.2822613.

55. Falk, A., Becker, A., Dohmen, T. J., Huffman, D. & Sunde, U. The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences. (2016) doi:10.2139/ssrn.2725874.

56. Bordalo, P., Gennaioli, N. & Shleifer, A. Memory, Attention, and Choice. (2017)

doi:10.3386/w23256.

57. Herzog, S. M. & Hertwig, R. The wisdom of many in one mind: improving individual judgments with dialectical bootstrapping. *Psychol. Sci.* **20**, 231–237 (2009).
58. Rolison, J. J., Hanoch, Y. & Freund, A. M. Perception of risk for older adults: differences in evaluations for self versus others and across risk domains. *Gerontology* 1–13 (2018) doi:10.1159/000494352.
59. Engel, C., Fedorets, A. & Gorelkina, O. How Do Households Allocate Risk? *MPI Collective Goods Discussion Paper* **14**, (2018).
60. Rohrer, J. M., Egloff, B., Kosinski, M., Stillwell, D. & Schmukle, S. C. In your eyes only? Discrepancies and agreement between self- and other-reports of personality from age 14 to 29. *J. Pers. Soc. Psychol.* **115**, 304–320 (2018).
61. Dohmen, T., Quercia, S. & Willrodt, J. Willingness to take risk: The role of risk conception and optimism. *SOEPpapers* (2019).
62. Rolison, J. J. & Shenton, J. How much risk can you stomach? Individual differences in the tolerance of perceived risk across gender and risk domain. *J. Behav. Decis. Mak.* **14**, 1085 (2019).
63. Wood, A. M., Brown, G. D. A., Maltby, J. & Watkinson, P. How are personality judgments made? A cognitive model of reference group effects, personality scale responses, and behavioral reactions. *J. Pers.* **80**, 1275–1311 (2012).
64. Schild, C., Ścigała, K. & Zettler, I. Reference Group Effect. in *Encyclopedia of Personality and Individual Differences* 1–3 (Springer, Cham, 2018). doi:10.1007/978-3-319-28099-8_840-1.
65. Menon, G., Raghubir, P. & Schwarz, N. Behavioral Frequency Judgments: An Accessibility-Diagnosticity Framework. *J. Consum. Res.* **22**, 212–228 (1995).

66. Blair, E. & Burton, S. Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Questions. *J. Consum. Res.* **14**, 280–288 (1987).