Supplemental Material: Through the Window of My Mind: Mapping the Cognitive Processes Underlying Self-Reported Risk Preferences

Markus D. Steiner
University of Basel

Florian I. Seitz
University of Basel

Renato Frey
University of Basel and Princeton University

Contents

Inclusion Criteria

Only participants who had completed at least 500 tasks on MTurk with an approval rate of at least 95% where eligible for our studies. Participants had to pass two out of two instructional manipulation checks (IMCs) 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. Moreover, when inspecting the aspects listed in the pilot study (see preregistration), we found that some participants either just entered a few letters or entered the words *no*, *nope*, or *none*. We thus decided to also exclude all aspects with fewer than four characters, or that consisted only of the words *nope* or *none*. Finally, participants had to complete the study either on a desktop computer or a laptop. In Study 1, we collected data of new participants until we had data of 250 participants that fulfilled all these inclusion criteria. In Study 2, the same criteria applied. Of the 164 participants who completed Study 2, we had to exclude ten because they did not meet the specified criteria. More specifically, two participants failed an IMC, one participant reported a focus of less than 25, and seven participants reported to not have completed the study on either a desktop computer or a laptop. Thus, our final sample size for Study 2 was 154.

Sentiment Analysis

To obtain the aspects' sentiments, we used the *tidytext* R package (Silge & Robinson, 2016), which provides different methods of scoring the sentiment of words. To this end, we first removed stop words (e.g., *and*), and then explored different scoring methods to find out which method works best given the resulting data structure. More specifically, we first tried the *afinn* lexicon which provides sentiment ratings of words between -5 and 5. However, as it has only a relatively small word lexicon, we could not obtain a sentiment for 13.9% of participants and thus there were many missing values. Because of this, and in line with our preregistration, we used the *bing* lexicon, which only provides binary sentiment ratings of words (positive or negative), but has a much larger lexicon. For every aspect, we then counted the positive and negative words and used the difference as sentiment score.

Deviations from Preregistered Analysis Plan

In this section we describe where and why we deviated from our preregistered analysis plan. All deviations were relatively minor and occurred in the analysis of Study 2.

First, we preregistered that the similarity ratings would be conducted by three independent raters who would rate the similarity of all possible aspect pairs (4286). To complete this work the raters would have been able to split the ratings over a maximum of four consecutive days. We recruited workers on Amazon MTurk for this task, but it turned out that many of them dropped out. We thus decided to split the workload and split the 4286 aspect pairs in 21 groups of about 200 aspects, which were each rated by three raters. We thus obtained data of 63 workers (instead of only three, as originally planned). The workers had to pass at least 6 out of 8 IMCs to ensure a high data quality.

Second, we adapted the way we calculated the proportion of overlapping aspects. We preregistered that the number of aspects listed in Study 1 would be the upper bound of possible overlapping aspects (i.e., the denominator in the fraction to obtain the proportion of overlap) and that two aspects would be considered to overlap if the mean similarity rating was at least four. As this approach led to values exceeding one, we adapted the way of obtaining the proportion of overlap in two respects. First, we set the threshold to be more conservative, more specifically, to five. Second, we used the smaller number of the number of aspects listed in Study 1 and Study 2, respectively, as denominator.

Third, we preregistered that we would use a linear model (with gaussian distribution and identity link function) to test the relation between aspect stability and the stability of self-reported risk preferences. This model identified no credible association (b = -1.21, , 95% CI: [-2.79, 0.36]). However, as both the outcome and the predictor variable were absolute difference scores, the errors were not normally distributed and the model thus likely biased. Therefore, we decided to run (and report in the main text) a gamma regression (i.e., with gamma distribution and log-link function). Note that this analysis found no credible association between aspect stability and the stability of self-reported risk preferences either. Relatedly, we explored additional ways of aggregating the similarity ratings across raters. These are reported in the additional analyses section below.

Fourth and finally, we preregistered that we would test the association between evidence stability and the stability of self-reported risk preferences with a linear model using the absolute difference scores between Study 1 and Study 2 of the two variables. With this model, evidence stability was a credible predictor of the stability of the reported risk preferences ($b = 0.04$, 95% CI [0.03, 0.05], $r_s = .41$). However, because only a directional test (i.e., without taking the absolute value of the differences) yielded normally distributed errors of the model, we also ran a directional Bayesian linear model in which, for better interpretability, we also z-standardized both difference scores. Using this model, we again found a credible association between evidence stability and the stability of the reported risk preference ($\beta = 0.63$, 95% CI [0.50, 0.75], $r_s = .45$). We decided to report the second model as some of the necessary assumptions in the first model were not met.

Details on Reported Analyses

Estimated Parameters for Query Theory

The positive regression coefficient for the SMRD in query theory (QT; averaged over the five coefficients obtained from the cross-validation procedure, mean $b = 1.76$, 95% CI: [1.36, 2.17]) indicated that there was a clustering of pro- and contra-aspects, respectively. Moreover, 90.4% of the SMRDs were either one or negative one, and 82.4% of the participants listed only pro- or only contra-aspects. Thus, this clustering was present very strongly. Furthermore, in line with previous findings (Johnson, Häubl, & Keinan, 2007), there was a positive correlation between participants' SMRD and their self-reported risk preference ($r_s = .80$), indicating that participants with a higher risk preferences tended to first report pro-aspects. The other parameters of QT indicated that the number of pro-aspects listed by a participant was credibly associated with higher risk-taking propensity ratings (mean $b = 0.49$, 95% CI: [0.31, 0.68]), and that the number of contra-aspects listed by a participant was credibly associated with lower risk-taking propensity ratings (mean b = -0.25, 95% CI: [-0.41, -0.08]), indicating that also the weight of evidence is a valid predictor of risk preference (but note that the weight of evidence and the strength of evidence correlate strongly, as the former is a binarized version of the latter).

Mixed Models Used to Quantify the Differences Between Pro- and Contra-Aspects

To quantify the effects of pro- versus contra-aspects, we ran generalized linear mixed effects models with the content variables (e.g., the rating whether an aspect involved an active choice or a passive experience) as outcome variables, and aspect valence (pro or contra) as dummy coded predictor. Additionally, we included by-subjects random slopes and intercepts, that is, we specified the maximal model (see, Barr, Levy, Scheepers, & Tily, 2013). Note that these analyses were not preregistered. We found that the estimated probability for an aspect describing a personal experience was credibly higher for the pro-aspects (p = .854) than the contra-aspects (p = .474; b = 1.87, 95% CI: [0.82, 3.17]). Moreover, the estimated probability that an aspect would involve an active choice rather than a passive experience was higher for pro- (p = .913) than for contra-aspects (p = .713; b = 1.44, 95% CI: [0.81, 2.22]). For the other two comparisons we found no credible effects of pro- versus contra-aspects (social comparison: b = 0.61, 95% CI: [-0.61, 1.91]; controllable vs. uncontrollable: b = 0.24, 95% CI: [-0.52, 1.08]). To quantify the difference in the reported frequencies between pro- an contra-aspects, we ran an ordinal mixed effects model using the *brms* R package (Bürkner, 2017), with the frequency categories as outcome variable, and aspect valence as dummy coded predictor. We again included by-subjects random slopes and intercepts. There was no credible effect of pro- or contra-aspects on the frequency categories (b = 0.28, 95% CI: [-0.24, 0.81]).

To also quantify the differences in the sentiment for pro- and contra-aspects, we then ran a linear mixed effects model with the sentiment as outcome variable and the aspect valence as dummy coded predictor. Moreover, we included by-subjects random slopes and random intercepts. The analysis showed a credible effect of the aspect valence indicating that pro-aspects, on average, had a higher sentiment than contra-aspects (b = 0.60, 95% CI: [0.40, 0.78]).

Priors Used in the Regression Models

Here we report the priors used in our Bayesian regression analyses. In all regression analyses we relied on the default priors of *rstanarm* and *brms* (Bürkner, 2017; Goodrich, Gabry, Ali, & Brilleman, 2018). That is, for all linear regressions we used, the prior for the intercept was $\mathcal{N}(0, 10)$, for the coefficients the prior was $\mathcal{N}(0, 2.5)$. In the case of the models with a gamma distribution and log link function there is an additional shape parameter that had the prior $exp(1)$. For the additional parameters of the (ordinal) mixed models, we refer to the R code for the detailed priors (also here we used the default priors).

Robustness Check of Modeling Results Using Participants Who Listed More Than One Aspect

In the case where only one aspect is listed, many of the models (SUM, FIRST, EXT, and LAST) make identical predictions. Thus, to test the robustness of our modeling analysis, we reran the cross-validation procedure of Study 1 but including only the data of participants who had listed at least two aspects (N = 223). The results of this robustness check were very similar to the original analysis with all participants, that is, the VUM still clearly outperformed the second best model (see Figure S1). Also, the mean ϕ parameter of the VUM was robust in that it still indicated a recency effect.

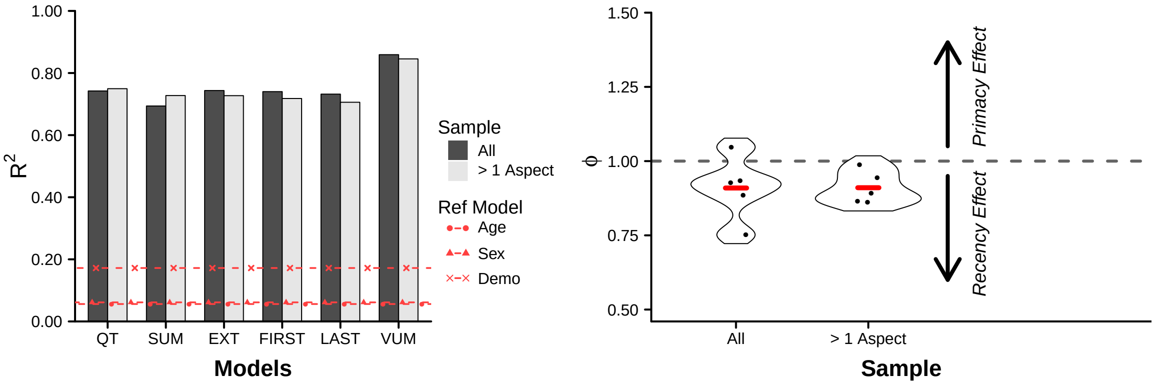

Figure S1. R^2 values of the different models predicting self-reported risk-taking propensity (left panel), and the ϕ parameters estimated for the VUM, as obtained from a five-fold cross-validation procedure (right panel). QT = Query theory; SUM = Sum of evidence; EXT = Most extreme evidence; FIRST = First aspect; LAST = Last aspect; VUM = Value updating model; R^2 = Variance explained by the model (in independent holdout-sets). "All" = Data of all participants were included in the analysis. "> 1 Aspect" = Only data of participants who listed more than one aspect were included in the analysis (N = 223). "Age" / "Sex" = R^2s of the reference models predicting risk preferences with age and sex as a single predictor, respectively. "Demo" = R^2 of the reference model predicting risk preferences with age, sex, years of education, income, and employment status as predictors. Red, horizontal lines in the right panel indicate the mean ϕ parameter value obtained in the five-fold cross validation procedure.

Robustness Check of Study 1

To test the robustness of the findings of Study 1, we reran the analyses of Study 1 with the data of Study 2. In the aspect listing, again, a majority of 121 participants (78.6%) listed between one and four aspects (M = 3.58; range: 1 - 13). Of these listed aspects, again 63% were contra-aspects. Moreover, most participants (N = 119; 77.3%) listed either only contra-aspects or only pro-aspects, and only a minority listed aspects of both types. Through this, the strength of evidence ratings within participants were rather stable with an intraclass correlation of .71.

The cognitive models were again very successful and the VUM again outperformed the second best model (FIRST) in terms of R^2 by 10 percentage points (see Figure S2). However, the R^2 values of the models were lower compared to those obtained in Study 1 (but still very high, between .57 and .71). To examine the predictive accuracy of the models over a longer period, we additionally fitted QT and VUM on the whole data set of Study 1 and used these parameters to predict the whole data set of Study 2. Again the VUM had the highest R^2 with .70 and again outperformed the second best model (QT) by almost 10 percentage points. Also regarding the contents of the listed aspects, the proportions were rather similar to those reported in Study 1, with the same patterns of credibility as in Study 1, except that now the ordinal mixed effects model in Study 2 revealed a credible effect of pro-aspects vs contra-aspects on the probability of the frequency category allotted to a reported aspect ($b = 0.96$, 95% CI: [0.32, 1.64]). Finally, pro-aspects, on average, again had a higher (i.e., less negative) sentiment than contra-aspects ($b = 0.44$, 95% CI: [0.20, 0.68]).

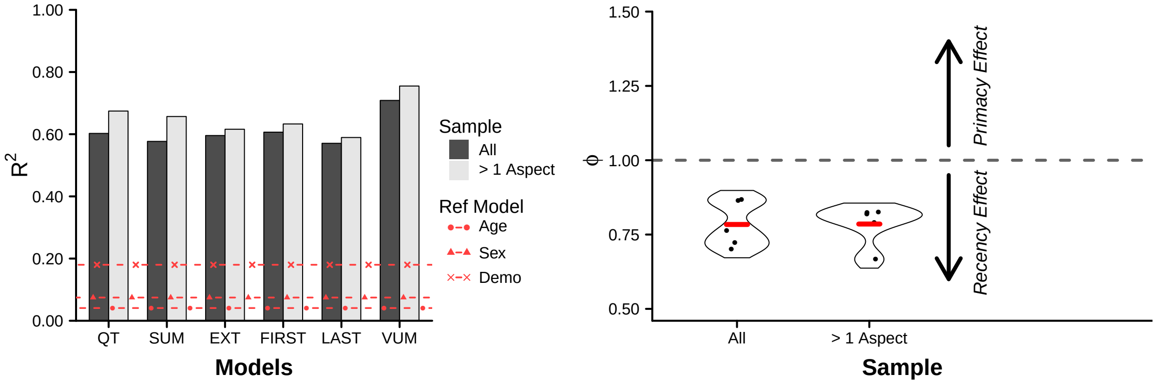

Figure S2. R^2 values of the different models predicting self-reported risk-taking propensity (left panel), and the ϕ parameters estimated for the VUM, as obtained from a five fold cross-validation procedure (right panel). QT = Query theory; SUM = Sum of evidence; EXT = Most extreme evidence; FIRST = First aspect; LAST = Last aspect; VUM = Value updating model; R^2 = Variance explained by the model (in independent holdout-sets). "All" = Data of all participants was included in the analysis. "> 1 Aspect" = Only data of participants who listed more than one aspect was included in the analysis (N = 137). "Age" / "Sex" = R^2s of the reference models predicting risk preferences with age and sex as a single predictor, respectively. "Demo" = R^2 of the reference model predicting risk preferences with age, sex, years of education, income, and employment status as predictors. Red, horizontal lines in the right panel indicate the mean ϕ parameter value obtained in the five-fold cross validation procedure.

Additional Analyses

Distribution of Self-Reported Risk Preferences

There were signs for a bimodal distribution of self-reported risk preferences, in line with observations of previous investigations (Dohmen et al., 2011; Frey, Pedroni, Mata, Rieskamp, & Hertwig, 2017), yet slightly stronger so (see Figure S3). A possible explanation for this amplification could lie in the deliberate reasoning of participants in our study: The explicit listing of (positive and negative) aspects may have led to slightly more extreme ratings. Furthermore, other patterns from the past literature could also be replicated in our dataset, such as that risk preferences were negatively associated with age (r_s = -.23 in Study 1, and r_s = -.19 in Study 2; Josef et al., 2016; Mamerow, Frey, & Mata, 2016.

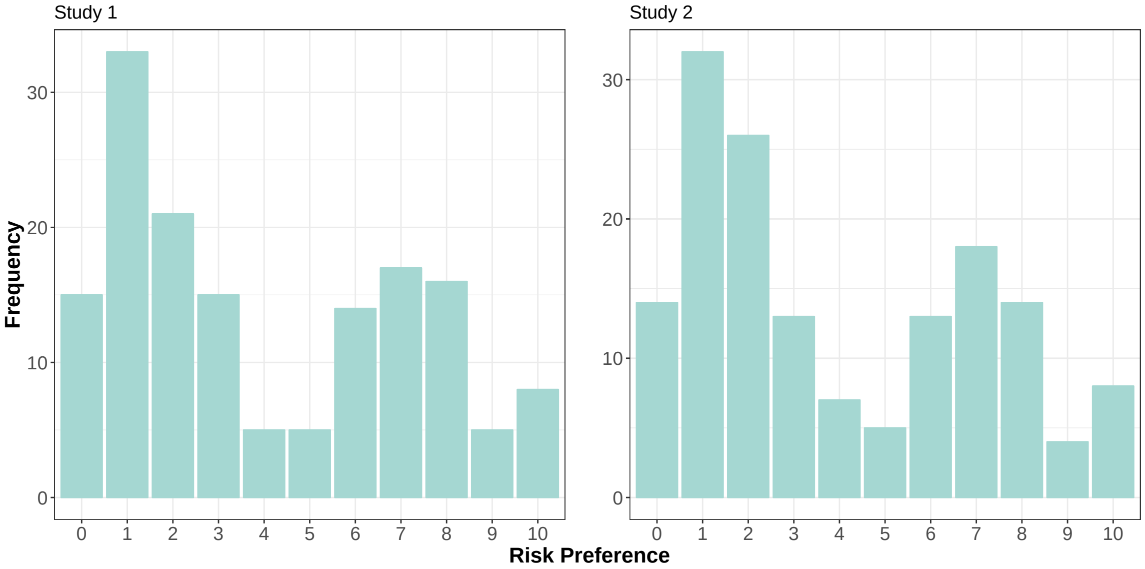

Figure S3. Distributions of self-reported risk preference in Study 1 (left) and Study 2 (right).

Additional Ways to Aggregate Similarity Ratings

We ran additional robustness checks on the relation between aspect stability and the stability of self-reported risk preferences. Specifically, we computed the similarities in two additional ways, (a) by taking the median instead of the mean, and (b) by identifying the rater pair with the highest Kendall's W of all three combinations of each rater triplet, and by then computing the average similarity only from the ratings of these two raters.

When we used the median to aggregate the similarity ratings, the average similarities within and between studies were very similar, though slightly higher compared to those reported in the main text (between studies: $M = 2.80$; within Study 1: $M = 2.81$; within Study 2: $M = 2.74$). From these similarity ratings, we then also computed the proportion of overlap. On average, we again found only a rather small proportion of overlaps (proportion of overlap across studies: .09; within Study 1: .05; within Study 2: .04). Using these overlap measures, we found no credible association between the absolute difference in the risk-taking propensity between Study 1 and Study 2, and the proportion of overlap between Study 1 and Study 2 in a participant's listed aspects using a GLM with a gamma distribution and a log link function (b = -1.15, 95% CI [-2.88, 1.06]).

When we used the mean of the ratings of the two raters with the highest Kendall's W to aggregate the similarity ratings, the average similarities within and between studies were again very similar, though this time slightly lower compared to those reported in the article (between studies: $M = 2.67$; within Study 1: $M = 2.72$; within Study 2: $M = 2.64$). From these similarity ratings, we then again computed the proportion of overlap. On average, we again found only a rather small proportion of overlaps (proportion of overlap across studies: .07; within Study 1: .03; within Study 2: .02). Using these overlap measures, we again did not find a credible association between the absolute difference in the risk-taking propensity between Study 1 and Study 2, and the proportion of overlap between Study

1 and Study 2 in a participant's listed aspects using a GLM with a gamma distribution and a log-link function (*b* = -2.52, 95% CI [-4.99, 0.40]). In sum, whereas we found clear associations between evidence stability and the stability of self-reported risk preferences, these robust checks indicate that further investigations are needed to clarify the impact of aspect stability.

References

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

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

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

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

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)

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

Josef, A. K., Richter, D., Samanez-Larkin, G. R., Wagner, G. G., Hertwig, R., & Mata, R. (2016). Stability and change in risk-taking propensity across the adult life span. *Journal of Personality and Social Psychology*, *111*, 430–450. doi: 10.1037/pspp0000090

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

Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in r. *The Journal of Open Source Software*, *1*(3), 1–3. doi: 10.21105/joss.00037