Comparing the Temporal Stability and Convergent Validity of Risk Preference Measures: A Meta-Analytic Approach

Alexandra Bagaïni[1], Yunrui Liu[1], Madlaina Kapoor[1], Gayoung Son[2], Paul-Christian Bürkner[3], Loreen Tisdall[1], and Rui Mata[1]

[1]Center for Cognitive and Decision Sciences, University of Basel

[2]Department of Psychology, University of Bern

[3]Department of Statistics, TU Dortmund University

Version date: July 11, 2023

Author Note

Correspondence concerning this article should be addressed to Alexandra Bagaïni, Center for Cognitive and Decision Sciences, Faculty of Psychology, University of Basel, Missionsstrasse 60-62, 4055 Basel, Switzerland, E-mail: alexandra.bagaini@unibas.ch

Abstract

Understanding whether risk preference represents a stable, coherent trait is central to efforts aimed at explaining, predicting, and preventing risk-related behaviours. We help characterise the nature of the construct by adopting a meta-analytic approach to summarise the temporal stability and convergent validity of over 300 risk preference measures (51 samples, 29 panels, >500.000 respondents). Our findings reveal significant heterogeneity across and within measure categories (propensity, frequency, behaviour), domains (e.g., investment, occupational, alcohol consumption), and sample characteristics (e.g., age). Specifically, while self-reported propensity and frequency measures of risk preference show a higher degree of stability relative to behavioural measures, these patterns are moderated by domain and age. Crucially, an analysis of convergent validity reveals a low agreement across measures, questioning the idea that they capture the same underlying trait. Our results raise concerns about the coherence and measurement of the risk preference construct.

keywords: risk preference, test-retest, age differences, life span

Comparing the Temporal Stability and Convergent Validity of Risk Preference Measures: A Meta-Analytic Approach

Risk permeates all domains and stages of life. Consequently, preferences towards risk may fundamentally shape individuals' health, wealth, and happiness. Risk preference—an umbrella term used to reflect the individual's appetite for risk (Mata et al., 2018; Schonberg et al., 2011)—not only has been related to personal decisions (e.g., timing of marriage and parenthood; Schmidt, 2008), but may also be used as an indicator to match individuals with products, services, and suitable careers (Breivik et al., 2019; Caliendo et al., 2014; Financial Services Authority, 2011; Jin et al., 2020). Because of its broad significance, risk preference is central to many theories and applications in the behavioural sciences (Barseghyan et al., 2018; Steinberg, 2013).

Despite the importance of the construct, there is considerable discussion about its central characteristics, including whether risk preference represents a stable, coherent trait or rather a contextual and/or domain-specific disposition (Schildberg-Hörisch, 2018; Schonberg et al., 2011; Stigler & Becker, 1977). One crucial source of the confusion surrounding the nature of risk preference is the many ways it has been operationalised. Specifically, the assessment of risk preference spans three measurement traditions that can be classified into three broad categories of measures: propensity, frequency, and behavioural measures, which, in turn, can differ in the domain (e.g., health, financial) and mode of assessment (e.g., ratings, choices; cf. Table 1). Crucially, past work suggests that different measures do not speak with one voice (e.g., Frey et al., 2017; Schonberg et al., 2011). As a consequence, resolving the debate about whether risk preference shares two central characteristics of a trait, namely stability and coherence, cannot be done without acknowledging the central role of measurement. Standing in the way of clarity, however, is the piecemeal approach taken in much past research, whereby single or few measures are adopted in any given study, making it difficult to obtain an overview across measures. Our work aims to help resolve this issue by taking a meta-analytic approach to investigate both the temporal stability and convergent validity of extant measures of risk preference.

A first focus of our work is to quantify the temporal stability of risk preference measures. This goal aligns with a key objective of discerning the sources of stability and change in human psychology and behaviour (Fraley & Roberts, 2005), and mirrors existing research into other traits (Anusic & Schimmack, 2016; Bleidorn et al., 2022; Elliott et al., 2020; Enkavi et al., 2019). Although some studies in economics and psychology have already probed the temporal stability of risk preference (e.g., Chuang & Schechter, 2015; Mata et al., 2018; Schildberg-Hörisch, 2018), there is a lack of a comprehensive comparison across measures with at least three significant gaps in existing research. First, previous work found higher stability for propensity and frequency measures than behavioural measures (Frey et al., 2017; Mata et al., 2018) but did not fully consider the role of domain (e.g., health, financial; Mata et al., 2018), leading to an oversimplified picture of the stability of measures. Second, there is little consideration of how the stability of different psychological constructs varies across the lifespan (Anusic & Schimmack, 2016; Bleidorn et al., 2022). Early life and young adulthood, which are marked by significant biological, cognitive, and social changes, usually show lower rank-order stability (Seifert et al., 2022) but past syntheses of the stability of risk preference did not account for age differences (e.g., Chuang & Schechter, 2015; Mata et al., 2018). Third, previous research has not employed theoretically grounded models to analyse temporal stability patterns across different categories of measures, domains, or populations, hindering comparison with other constructs, such as major personality traits, that have been studied using formal models (Anusic & Schimmack, 2016).

A second focus of our work is to quantify the convergent validity of risk preference measures. The issue of convergence is central to the goal of mapping theoretical constructs to specific measures and many efforts in the behavioural sciences aim to empirically estimate these links (Duckworth & Kern, 2011; Eisenberg et al., 2019; Frey et al., 2017). The issue is also of practical importance because many studies investigating predictors or correlates of risk preference, for example, neuroimaging and genome-wide association studies (Karlsson Linnér et al., 2019; Karlsson Linnér et al.,

2021), are often able to use only a single or limited set of measures to capture risk preference. To the extent that different measures do not speak with one voice, however, these should not be used interchangeably and need to be carefully selected to match the construct of interest. Previous work on risk preference reports a relatively low convergence between measures, albeit propensity and frequency measures may exhibit moderate convergent validity among themselves, whereas behavioural measures show comparatively low convergent validity, in terms of both observable behaviour and computational parameters (Frey et al., 2017; Pedroni et al., 2017). We note three key gaps in extant work on the convergent validity of risk preference measures. First, studies typically employ only a few different measures, thus limiting the extent to which a comprehensive assessment of convergence between many measures can be performed in a single study. Second, the adoption of few measures in single studies often implies that the influence of measure (e.g., category, domain) or respondent characteristics (e.g., age) cannot be ascertained as moderating variables that can impact the convergence of measures. Third, and finally, studies have not been able to assess the extent to which low convergent validity is a direct result of poor reliability of specific measures (Dang et al., 2020; Strickland & Johnson, 2021).

The present study tackles these outstanding gaps by examining the temporal stability and convergent validity of a comprehensive set of risk preference measures. For this purpose, we conducted a systematic search for longitudinal data sets comprising many different measures of risk preference, including propensity, frequency, and behavioural measures. The curated database represents a large data trove comprising 29 longitudinal panels, split into 51 different samples, capturing over 300 different measures of risk preference. To further enhance the comprehensiveness of this newly curated data, we conducted an extensive categorisation of measures (e.g., category, domain) and associated respondents (e.g., age, gender).

Equipped with these data, we conducted a number of analyses to gain an overview of the temporal stability and convergent validity of risk preference measures. First, to comprehensively examine temporal stability, we performed a variance

decomposition analysis that provides a picture of the amount of variance that can be accounted for in temporal stability by measure, respondent, and panel-related predictors. We also adopted a formal modelling approach using the meta-analytic stability and change model (MASC; Anusic & Schimmack, 2016) to capture the temporal stability of risk preference measures while distinguishing between domains (e.g., investment, gambling, smoking, ethical). We further employed MASC to re-analyse longitudinal panel data for other pertinent psychological constructs, including personality and affect, thus providing a direct comparison between our results and those for other major psychological constructs. Second, to comprehensively examine convergent validity, we performed variance decomposition analysis to quantify to what extent measure, respondent, and panel-related predictors account for the heterogeneity observed between inter-correlations. Crucially, because it has been suggested that the reliability of individual measures creates boundary conditions for their convergence (Dang et al., 2020), we consider measure reliability as a measure-related predictor in these analyses. We further report meta-analytic syntheses of the empirical relation across measures both between and within category and domain pairs. All in all, we hope that by clarifying the two central characteristics of measures of risk preference—temporal stability and convergent validity— we can contribute to improving its measurement, describing its life course patterns, and, ultimately, its utility as a construct in the behavioural sciences.

Results

Overview of the Longitudinal Data

We used a systematic approach to identify a comprehensive set of longitudinal samples suitable for estimating the temporal stability and convergent validity of risk preference measures. Figure 1 depicts the flow of steps starting from the identification of panels, screening for eligibility, and, finally, the data available for the temporal stability and convergent validity analyses. Please note that we distinguish between

panels and samples because if panels included data from several countries, we treated these as separate samples to avoid confounding within-and cross-country differences. As per our inclusion criteria, all the samples had to contain at least one propensity measure. This criterion was implemented to enable comparisons between propensity measures, the most prevalent category in the literature on risk preference, to other categories (i.e., frequency, behaviour) as well as to similar measurement approaches in personality research (cf. Anusic & Schimmack, 2016). From the initial pool of 101 panels (157 samples) identified in our search, we were able to include 29 panels (51 samples) that allowed computing test-retest information for at least one measure of risk preference, and 26 panels (45 samples) that allowed computing inter-correlations between two or more measures of risk preference. Finally, for each risk preference measure, sample, age group, and gender, we calculated test-retest correlations between all measurement wave combinations for temporal stability analyses, and all possible inter-correlations between measures for convergent validity analysis. This process yielded over 72,000 test-retest correlation coefficients for temporal stability (Figure 2A) and over 61,000 inter-correlations for convergent validity analyses (Figure 2B). As a whole, the dataset covers over 300 different measures of risk preference spanning three measure categories (i.e., propensity, frequency, behaviour).

Informed by previous work that has distinguished between different domains of risk, we conducted an extensive categorisation of measures to distinguish between 14 different domains (e.g., general health, financial, recreational, driving), thus allowing a fine-grained classification sorely lacking in the risk preference literature. Crucially, this categorisation makes clear that there are important differences across, and also gaps between, the domains investigated in each category. As can be seen in Figure 2C, while propensity measures capture the majority, albeit not all, of the domains detected in our data (9 out of 14), frequency measures capture a large but different subset of these (8 out of 14). In turn, behavioural measures capture only a small minority of finance-related domains, such as investment and gambling (4 out of 14). This imbalance is ultimately due to the different traditions spanning the psychology, economics, and

public health literature that have investigated risk preference using different measurement categories. In what follows, we provide a fine-grained comparison of the measures' temporal stability.

Temporal Stability

We first obtained an overview of the temporal stability data by visualising the number of measures by category and retest interval as well as a breakdown of the test-retest correlations by measure category (propensity, frequency, behaviour; see Figure S1A). We should note that there are substantial differences in the amount of data concerning different categories, with most measures being classified as propensity or frequency measures and only a minority as behavioural measures. The under-representation and overall shorter test-retest intervals for behavioural measures observed in our sample is a product of there being overall fewer samples that have included (repeatedly) such measures in their assessment batteries, likely due to the additional burden of deploying behavioural measures which typically require extensive instructions, multiple choices, and, potentially, incentivisation. Figure S1B provides a first impression of the distributions of retest correlations across time and measure categories that conveys considerable heterogeneity between measures that we explore quantitatively in more detail below.

Variance Decomposition of Test-Retest Correlations

Our first main question concerns the relative contribution of measure, respondent, and panel characteristics in accounting for patterns of temporal stability in different measures of risk preference. For this purpose, we adopted a Shapley decomposition approach, a method that estimates the average marginal contribution of different predictors to the variance in an outcome of interest (Grömping, 2007), in our case, the test-retest correlations of risk preference measures. We were particularly interested in the role of specific measure- and respondent-related predictors that have been either hypothesised or shown to account for some variance in temporal stability in past work on risk preference (e.g. Frey et al., 2017; Josef et al., 2016) or other

psychological constructs (e.g. Anusic & Schimmack, 2016). For measure-related predictors, we focused on the category (i.e., propensity, frequency, behaviour), domain (e.g., general health, recreational), the scale type (e.g., ordinal, open-ended), and length of the test-retest interval (e.g., 6 months, 1 year, 5 years). For respondent-related predictors, we considered age group, gender, and number of respondents. Finally, we also included panel as a predictor to capture the role of unobserved panel characteristics (e.g., quality of data collection or data entry) that can influence test-retest reliability.

We first conducted an omnibus analysis to assess to what extent measure, respondent, and panel predictors explained differences across all test-retest correlations. Altogether, a model considering all predictors captures 49.7% of the observed variance. As can be seen in Figure 3A, we find that a large portion of the variance could be explained by measure-related predictors, domain (13.7%), category (4.3%), retest interval (6.8%), and scale type (0.5%). In turn, we find that some of the variance could be explained by respondent-related predictors, in particular, age (5.2%). Finally, panel captured a large portion of the variance (18.7%), suggesting that there are a number of (unobserved) panel characteristics that also contribute to systematic differences in the observed temporal stability of measures.

Given our focus on comparing measure categories, we further explored the differences between the contribution of these predictors to propensity, frequency, and behavioural measures separately. The models conducted separately by measurement category explained 23.7%, 46.6%, and 16.6% of the total variance for propensity, frequency, and behavioural measures, respectively. The results of this analysis are depicted in Figure 3B. There are four main insights that can be drawn from the comparison between measure categories. First, domain explained a significant percentage of the variance for frequency (12.5%) relative to propensity (1.3%) and behavioural (5.6%) measures. This suggests considerable heterogeneity within some categories as a function of domain, in particular, in the frequency category, something we will explore in more detail when analysing the temporal trajectories by domain below. Second, retest interval contributed to more explanatory power for propensity

(5.2%) and frequency (6.9%) measures relative to behavioural measures (1.0%), suggesting that the temporal patterns are less pronounced for the latter. Third, concerning respondent-related predictors, we find that age explained a significant percentage of the variance in the test-retest correlations, but, in particular, for frequency (8.4%) relative to propensity (2.3%) and behavioural (0.8%) measures. These results seem to indicate some specificity regarding the effects of age by measure category. Fourth, as in the omnibus analysis, a number of (unobserved) panel characteristics seem to contribute to systematic differences between panels, albeit this effect is most pronounced for frequency measures. In what follows, we explore these results in more detail by adopting a formal modelling approach that distinguished between the different measure categories and domains.

Meta-Analyses of Temporal Stability

We used the Meta-Analytic Stability and Change model (MASC; Anusic & Schimmack, 2016) to capture the trajectory of test-retest correlations across measures of risk preference and compare these to other psychological constructs. MASC uses three parameters to represent different properties of temporal trajectories: reliability (proportion of between-person variance excluding random error), change (proportion of variance that is subject to changing factors), and stability of change (the rate at which change occurs over time). In our work, we adopted a sampling-based Bayesian estimation procedure to obtain full posterior distributions for each model parameter for specific measure categories (propensity, frequency, behaviour) and domains (e.g., recreational, general health, smoking, investment).

Figure 4 shows model predictions for the trajectory of test-retest correlations separately for the three measure categories and distinguishing further between domains (e.g., recreational, general health, smoking, investment) and respondent groups (age groups, gender). Figures 4A-C show the distributions of the predictions for each of the model parameters, while Figures 4D-I show the corresponding trajectories in test-retest correlations as a function of retest interval for different age groups (panels D, F, H), as well as the (equivalent) age trajectories as a function of different retest intervals (panels

E, G, I). While the trajectories in test-retest correlations as a function of retest interval are particularly helpful to compare to similar trajectories found for other psychological constructs (Anusic & Schimmack, 2016), the trajectories by age for different retest intervals help visualise a potential inverted U-shape function across the life span in patterns of reliability found in past work using propensity measures of risk preference (Josef et al., 2016) and major personality traits (Bleidorn et al., 2022).

We find a ranking in the overlapping reliability estimates for the three measure categories, with the highest reliability found for propensity measures (M: 0.51, 95% HDI: [0.42, 0.61]), followed by frequency measures (M: 0.47, 95% HDI: [0.33, 0.63]), and behavioural measures (M: 0.30, 95% HDI: [0.20, 0.40]). Crucially, relative to propensity and behavioural measures, the reliability of frequency measures varies widely by domain, with a wide range evident between the highest reliability for smoking (M: 0.84, 95% HDI: [0.78, 0.90]) and the lowest for the ethical domain (M: 0.11, 95% HDI: [0.04, 0.18]). In comparison, the ranges found for propensity measures, spanning from ethical (M: 0.64, 95% HDI: [0.36, 0.91]) to occupational (M: 0.41, 95% HDI: [0.32, 0.49]), and behavioural measures, spanning from investment (M: 0.36, 95% HDI: [0.24, 0.49]) to insurance (M: 0.26, 95% HDI: [0.17, 0.36]), are considerably smaller. Concerning the patterns of change and associated stability, the different measure categories and domains appear comparable and seem to mimic those found in the temporal stability literature characterised by steep changes yet some long-term stability (Anusic & Schimmack, 2016; Fraley & Roberts, 2005).

Concerning age-related patterns, we note clear trends for propensity and frequency measures but not behavioural ones. Specifically, as can be seen in Figure 4C, when considering longer retest intervals (>2 years) for propensity measures, and consistent with previous work (Josef et al., 2016), we note an inverse U-shape association between retest-correlations and age, indicating that temporal stability peaks in middle-age. Also, this pattern is observed for most domains covered by propensity measures (Figures S7-S9). For frequency, the overall pattern observed in Figure 4G is more mixed but we should note that this appears due to heterogeneity between domains

within the frequency category, as we observe an inverse-U shape with age for both alcohol consumption and smoking domains. In turn, the driving, ethical, and sexual intercourse domains do not show the same pattern (Figures S10-S11). For behavioural measures, as seen in Figure 4I, we do not observe noticeable association between temporal stability and age, and this is reflected across the individual domains (Figure S12). Concerning gender, we did not identify any substantial differences, suggesting males and females show comparable stability trajectories across the board.

Finally, we assessed where risk preference stands within the consistency hierarchy of psychological constructs (Conley, 1984), by comparing the temporal stability of risk preference to that of personality, life satisfaction, self-esteem, and affect using data of Anusic and Schimmack (2016). Our results obtained using a Bayesian framework largely replicate those of Anusic and Schimmack (cf. Figure S13) but allow us to compare directly the estimates for different constructs using the same modelling approach. Our reanalysis show highest reliability for personality traits (M: 0.73, 95% HDI: [0.68, 0.77]), followed by self-esteem (M: 0.62, 95% HDI: [0.54, 0.71]), life satisfaction (M: 0.60, 95% HDI: [0.55, 0.64]), and affect (M: 0.56, 95% HDI: [0.50, 0.61]). In line with the results for risk preference given above, this suggests that the average stability of risk preference as captured by propensity and frequency measures, is, on average, lower than that of major psychological constructs albeit it overlaps with that for affect. In turn, the reliability of behavioural measures is lower than any of the four constructs, suggesting a qualitative difference between this category and the constructs considered. Of course, as suggested above, for frequency measures, some domains show considerably higher/lower levels of stability; consequently, while frequency measures in the smoking and alcohol domains rival the temporal stability of major personality traits, others, like ethical and driving, show some of the lowest reliability estimates observed, suggesting these do not have the same stable quality.

All in all, the results on temporal stability support the notion that different risk preference measures show markedly different temporal stability signatures. In what follows, we explore further differences between measures by evaluating their

inter-correlations.

Convergent Validity

Variance Decomposition of Correlations Between Measures

We first obtained an overview of the convergent validity data by visualising the distributions of inter-correlations of measures separately for different measure pairs (Figure S14). The resulting pattern speaks to the large heterogeneity in correlations between measures as well as possible differences between and within measure categories. We used variance decomposition to provide a quantitative summary of correlations as a function of several measure and respondent-related characteristics, as well as panel. Specifically, concerning measure characteristics we included dummy-coded predictors to code for the matching (e.g., propensity-propensity) or mismatching category (e.g., propensity-frequency), domain, and scale type. Further, using the results from the temporal stability analyses above, we computed the average reliability of each pair of measures and included this in our predictors to assess the extent to which measures' reliability contribute to their convergence.

The variance decomposition analysis suggests that a model considering all predictors captures 26.6% of the variance in inter-correlations. More substantively, as shown in Figure 5, the variance decomposition analysis suggests that category and domain play a considerable role: More than half of the explained variance was accounted for by whether or not the pair of measures matched in terms of category (7.5%) and domain (10.0%). In turn, we find that measure reliability accounted for less than 1% of the variance, thus indicating little support for the idea that poor reliability of risk preference measures is the main driver of their (lack of) convergence. Finally, respondent-related effects offer little to no contribution, while panel characteristics seem to account for some amount of variance, suggesting that unobserved panel characteristics capture relevant, systematic variance in the correlation between measures. In sum, the variance decomposition analysis suggests that measure characteristics, specifically, category and domain, capture important aspects of measure

convergence. In what follows, we provide a more detailed overview of the role of these factors by providing a meta-analytic correlation matrix across pairs of measures that distinguishes between category and domain.

Meta-Analyses of Convergent Validity

We adopted a meta-analytic approach to map out the convergent validity of risk preference measures across categories and domains. For that purpose, we conducted separate meta-analyses at different levels of aggregation. A meta-analysis across all available inter-correlations, suggests an average meta-analytic inter-correlation of .16, 95% HDI: [0.13, 0.18]. However, this value hides considerable heterogeneity. As can be seen in Figure 6A, across pairs of categories and domains, we observe a large range of inter-correlations, from around -.2 to circa .8. The meta-analytic correlation matrix also shows evidence of overall higher average correlations along the diagonal, signalling that matching both category and domain leads to typically higher inter-correlations relative to matching only across domains or categories. Importantly, as can be seen in Figure 6B, when considering aggregation at the category level, there is a clear ranking of the average inter-correlations within category, with this being highest for propensity (M = 0.41, 95% HDI: [0.40, 0.43]), followed by frequency (M = 0.20, 95% HDI: [0.18, 0.22]), and behavioural measures (M = 0.19, 95% HDI: [0.15, 0.23]). Finally, and more importantly, there is evidence of little convergence between categories, with cross-category meta-analytic correlations being around or smaller than 0.1.

All in all, considering jointly the results on both temporal stability and convergent validity, one is left with the impression that different risk preference measures can show very different psychometric signatures, including patterns of temporal stability and convergent validity, supporting the notion that measurement issues plague clarity concerning the nature of the construct.

Discussion

Our aim was to contribute to the ongoing debate about whether risk preference represents a stable and coherent trait by adopting a meta-analytic approach to assess the temporal stability and convergent validity of a large set of risk preference measures. We curated an extensive collection of previously un(der)utilised longitudinal samples, providing data for over 300 unique measures rooted in measurement traditions that are aligned with the adoption of three broad categories of measures—propensity, frequency, and behavioural measures—and covering various life domains (e.g., driving, alcohol, smoking, social, ethical, recreational, occupational, gambling). Our work provides the first encompassing meta-analytic syntheses of the trajectories of temporal stability and convergent validity across these major measure categories while accounting for central measure (e.g., domain) and respondent (e.g., age) characteristics. Crucially, we do so by adopting a formal model of temporal stability that allows comparing the temporal stability trajectories results both between measures of risk preference as well as with other major psychological constructs.

Our analyses of the temporal stability of risk preference measures suggest some average differences in the reliability of measures from the three measurement traditions. Overall, propensity measures exhibited the highest average reliability, followed by frequency measures, while behavioural measures showed the lowest average reliability. Crucially, we observe considerable overlap between categories and substantial heterogeneity within categories as a function of domain. Particularly, and most profoundly for frequency measures, reliability varies widely between domains, with smoking showing the highest reliability, while others, such as the ethical domain (i.e., violent or delinquent behaviour) showing the lowest. Concerning respondent characteristics, we find that age affects the temporal stability patterns found for propensity and frequency measures, but not behavioural ones. Specifically, test-retest correlations were lower in younger and older age groups compared with middle-aged groups for propensity measures, a common pattern found for other personality constructs (Anusic & Schimmack, 2016; Bleidorn et al., 2022; Seifert et al., 2022). For

frequency measures, age patterns are more heterogeneous across domains. For example, while the smoking and alcohol consumption domains show increasing stability across adulthood followed by some decline in old age, the ethical and sexual domains shows patterns consistent with decreased stability in adolescence and young adulthood. We note that this heterogeneity maps onto distinct pathways for age-specific versus lifelong trajectories of these different behaviours (Ahun et al., 2020; Moffitt, 2018). The heterogeneity in the temporal patterns of risk preference measures poses a problem for its comparison with that of other psychological constructs. Nevertheless, one conclusion that emerges is that propensity measures show somewhat lower test-retest stability but similar age-related (inverted-U) trends compared to that of major personality traits. Frequency measures are more heterogeneous and, therefore, not easily compared as a whole, some domains, like smoking and alcohol consumption, approach the stability of personality constructs and show similar age patterns. In contrast, other domains captured by frequency measures, such as driving and ethical domains, show very low stability and most change occurring in adolescence and young adulthood. Behavioural measures show considerably lower stability compared to the other categories (i.e., propensity, frequency) or psychological constructs (e.g., self-esteem) and do not seem to capture any life span trends. As a whole, these results suggest that different measurement traditions are characterised by distinct temporal and age-related trajectories, emphasising the important role of measurement, domain, and age in moderating the patterns of temporal stability concerning risk preference measures.

Our analyses of convergent validity showed that, overall, convergence between risk preference measures was low, albeit highlighting substantial heterogeneity between measure categories. Convergence was highest for propensity measures, while frequency and behavioural ones showed lower convergence, somewhat matching previous results from individual studies (Eisenberg et al., 2019; Frey et al., 2017). One should note that frequency measures covered a considerably larger set of domains spanning health, occupational, and gambling domains compared to behavioural measures that shared a focus on financial domains (i.e., investment, gambling, insurance), which may present a

confound when estimating differences between these two categories. Unfortunately, frequency measures did not cover these financial domains well, making a direct comparison between frequency and behavioural measures impossible. Crucially, we found that relatively little variance in convergence between measures was explained by their average reliability, suggesting that there may be something more fundamental about measure characteristics that contributes to their lack of convergence. To sum up, somewhat mirroring the temporal stability analyses, the results on convergence suggest that different measurement traditions do not speak with one voice but, rather, show unique patterns by category and, particularly for frequency measures, are largely moderated by domain. In contrast with the temporal analyses, however, age did not seem to be a strong determinant of measure convergence. These results suggest the different measures cannot be used interchangeably to capture individual differences in risk preference and call into question the coherence of the risk preference construct.

Before we address the implications of our findings for our understanding of risk preference and its measurement, several limitations of our study should be noted concerning our 1) search and inclusion criteria, our 2) coding of predictor variables, and other 3) analytical choices. First, despite conducting an extensive search for panels, there may be additional ones that were missed by our independent research effort. Exploring yet more panels could lead to the discovery of additional measures that could further improve the scope of our findings. Further, our focus on comparison between measurement traditions as well as other psychological constructs led us to consider samples only if they included at least one propensity measure, which likely contributed to over-representation of this category relative to others (e.g., behavioural), as well as the domains represented across categories. A promising solution involves pursuing even more comprehensive efforts, for example by leveraging crowd sourcing or coordinated analyses across multiple research teams. By tapping into the collective expertise and resources of a broader community, one could make the efforts of mapping risk preference measures yet more exhaustive. Second, to assess the role of a set of theoretically relevant predictors for temporal stability and convergence, we meticulously coded

relevant information about the measures (e.g., category, domain, test-retest interval, scale type, pair type for convergence) as well as the respondents (e.g., age, gender). While we recognise the value of additional information (e.g., measure incentivisation, respondents' socioeconomic status), it proved challenging to obtain sufficient data to allow including more fine-grained comparisons in our analysis or ensure comparability across samples. Another coding issue concerns our use of panel as a predictor, which could have been broken down further (e.g., main data collection mode, language) but proved unfeasible to model in our framework. In light of these constraints, our coding scheme and analyses were geared towards including maximally informative predictors while ensuring computational feasibility. Perhaps future efforts including additional data can help resolve the role of additional moderating factors. Third, our workflow required making a number of analytical choices, including the binning of age groups, or the selection of statistical metrics and model priors in our Bayesian framework. Whenever possible, we made principled decisions informed by past work. To deal with this issue, we conducted multiverse analyses to assess the robustness of our results whenever possible. Finally, given the complexity of the data curation process we did not pre-register our analysis but we make our data and scripts publicly available which we hope will allow the research community to collaborate on future efforts to examine the psychometric characteristics of risk preference measures.

Our findings provide a new empirical overview on the status of many extant risk preference measures. We would like to point out four main implications of these findings for current theorising and empirical research on risk preference. First, our results indicate we need to invest new energy into developing theoretical frameworks that help us make sense of the observed convergence as well as divergence across measures. One factor leading to the gap between measures we have documented may arise from fundamentally different concepts of risk taking being captured by different measures (e.g., Bran & Vaidis, 2020). Specifically, propensity measures aim to capture individuals' attitudes towards risk, while frequency measures aim to capture actual risky behaviour, which will often be a product of both individuals' appetite for risk as

well as other considerations, including the opportunity to engage in these risks. In this sense, the gap observed between propensity and frequency measures could be interpreted as a special case of the classic intention-behaviour gap. This explanation, however, leaves the lower reliability of behavioural measures and their low convergence with propensity and frequency measures largely unresolved. Some researchers have pointed out current limitations of behavioural measures that can contribute to this state of affairs. For example, behavioural measures may require many trials to obtain reliable estimates of the underlying latent trait, something that is more easily and naturally accomplished by integrating behavioural episodes from memory (e.g., Haines et al., 2020). One other more general factor contributing to the gap between measures concerns the levels of granularity adopted. For example, while propensity measures are typically general, covering a broad domain (e.g., health) and time span ("in general"), frequency measures are more specific (e.g., "number of cigarettes") and constrained in time (e.g., "in the last 30 days"), and behavioural measures could perhaps be thought as yet more specific (e.g., about specific types of monetary choices). The lack of a direct match in levels of granularity can contribute to lower reliability because individuals may think of different aspects when answering general questions or even provide different answers depending on the cues that happen to come to mind on any given occasion (Arslan et al., 2020; Steiner et al., 2021). We would like to note that the effort to understand how these factors contribute to gaps between measures should not be seen as a simply methodological one. Clarifying the conceptual and empirical relations between constructs and how these are operationalised is central to achieving conceptual clarity in the behavioural sciences (e.g., Bringmann et al., 2022). Consequently, it should also be seen as part of a larger effort to integrate risk preference in the larger context of psychological constructs and associated ontologies (Eisenberg et al., 2019; Norris et al., 2019).

Second, in line with the focus on theory development, our results emphasise the need to understand the temporal stability of risk preference as a function of life span changes in a heterogeneous set of contexts or domains. Many extant theories make

valuable contributions to explaining the complex nature of stability and change in personality traits (Möttus et al., 2019) and behaviours, such as antisocial (Moffitt, 2018) or health behaviours (Ahun et al., 2020). Transactional models appear particularly promising in that they emphasise the interplay between individual characteristics and environmental factors in determining phenotypic change across the lifespan (Möttus et al., 2019). Our results suggest that such transactional models could be helpful in reconciling the idea of stable individual risk preferences with differential patterns across domains that are shaped by changing affordances and goals (Ravert et al., 2019) as well as individuals' life experiences (Beck & Jackson, 2022).

Third, from a more methodological perspective, our findings suggest it is important to streamline and replenish our methodological resources by focusing on principled measure validation and development. Regarding validation, we should strive for more comprehensive comparisons of existing measures. This can be achieved through meta-analytic research, similar to our current approach, as well as primary studies that explore previously overlooked measure categories, domains, and their combinations (Richmond-Rakerd et al., 2020). We also need to engage more actively with particular behaviours, and conduct targeted explorations of domains using multiple measures across different categories (cf. risky driving, Das & Ahmed, 2022). Regarding measure development, recent technological development suggests that there are new forms of measurement on the horizon that could help anchor measures of risk preference in more real-world experience, for example, through the use of virtual reality (Roberts et al., 2021), or text-based analysis facilitated by large language models (Wulff & Mata, 2022), as well as biology, through the use of advanced imaging methods that track structural aspects of neural processing of reward (Tisdall et al., 2022).

Fourth, and finally, we need to combine the improvements awaiting us in the development and validation of both theories and measures to focus on prediction. Three centuries ago, the topic of risk preference emerged from Daniel Bernoulli's interest in solving practical problems, aiming to use mathematical formalisation to help understand how individuals make consequential decisions regarding gambling, financial

investment, and insurance (Bernoulli, 1954). Principled prediction requires a good understanding of the anticipated mechanisms as well as an informed selection of measures (on the side of both outcomes and predictors). Future work will need to integrate objective measures in the domains of health (e.g., inflammation markers, visits to the emergency department), investment (e.g., stock portfolios), and ethics (e.g., arrest records, number of speeding tickets) to assess the predictive value of different risk preference measures. We hope this focus on prediction will ultimately fuel a better understanding of what risk preference means for whom and at what stage in their life thus buttressing the utility of the construct for predicting important life outcomes and ultimately improving individuals' health, wealth, and happiness.

Methods

Identification of Samples

We used a systematic method to find a comprehensive set of longitudinal data that include measures of risk preference (Figure 1). We started by identifying longitudinal panels by 1) performing searches on general-purpose search engines, survey listings, and data repositories (i.e., Google Database, Gateway to Global Aging Data, Gesis, IZA, ICPSR, CNEF, UK Data service) using relevant terms (e.g., "risk preference", "risk aversion", "risk attitude", "take risks", "survey", "panel", "longitudinal"; cf. Table S1 for a list of our search terms), 2) consulting past literature for references to longitudinal panels or studies that have estimated the temporal stability of psychological constructs (i.e., Anusic and Schimmack, 2016; Chuang and Schechter, 2015; Graham et al., 2020; Mata et al., 2018; Orth, 2018), and 3) informal requests to colleagues for suggestions concerning panels or specific studies. This search led to identifying 101 longitudinal panels (157 samples; Table S2). It is important to note that we differentiate between panels and samples, such that samples have their origin in a panel. For example, if a panel (e.g., SHARE) included data from multiple countries (e.g., SHARE-Switzerland, SHARE-Germany, SHARE-Belgium), we treated the latter as distinct samples to prevent confusion between differences within and across countries. To determine the relevance of each of the 157 samples for our analyses, we adopted a set of screening criteria (Table S3). In brief, we included a sample in our analyses if it 1) was publicly available, 2) included data on at least one consistently formatted propensity measure of risk preference with responses from the same respondents across at least two time points, and 3) included data on the gender and age of the respondents. This procedure led to the creation of a comprehensive data trove comprising 51 samples from 29 longitudinal panels (Table S4). For each sample, we included data that was available as of May 2023.

Categorisation of Measures

To further add to the comprehensiveness of the newly curated data set, we conducted a categorisation of each risk preference measure. The following measure characteristics are particularly relevant to our analysis: measure category (e.g., propensity, frequency, behaviour), domain (e.g., investment, general health, social, recreational), and scale type (e.g., open or closed questions). Table S5 presents descriptions of risk preference measures that are representative of the variety of measures included in the samples used for our analyses. With regards to the domains captured by different risk preference measures, we included measures covering as many domains as possible, that is, we did not exclude measures in pre-specified domains. Further, we adopted a bottom-up, data-driven approach mostly to distinguish between domains. We felt this approach was best suited for our purpose, as this allowed us to 1) scope extant work and systematically identify the domains most commonly assessed in the risk preference literature, and 2) provide the most comprehensive assessment to date of temporal stability and convergent validity while systematically investigating the role of domain at a high level of granularity. Overall, we identified 14 domains: *alcohol, driving, drugs, ethical, gambling, general health, general risk, insurance, investment, occupational, recreational, sexual intercourse, smoking, and social*. Our labelling scheme has considerable overlap with terminology commonly used to group contexts or situations within which risk taking can occur, albeit it makes fine-grained distinctions within domains, such as distinguishing between smoking or alcohol consumption from a more general health domain. We provide additional detail concerning an assessment of measure characteristics in the Supplementary Information.

Temporal Stability

In what follows, we give an overview of steps involved in computing test-retest correlations, conducting variance decomposition of test-retest correlations, and the modelling of temporal stability using the Meta-Analytic Stability and Change model (MASC; Anusic and Schimmack (2016)). We provide additional information concerning

each step in the Supplementary Information.

Computing Correlations

To compute test-retest correlations, we followed a similar approach as Anusic and Schimmack (2016) and Enkavi et al. (2019). For each panel we included the data from all the respondents, regardless of whether or not they provided responses on all measurement waves. Within each sample and for each risk preference measure, we calculated test-retest correlation coefficients for each possible wave combination. For example, for a sample with Waves 1, 2 and 3, we calculated three sets of test-retest correlations: between Wave 1 and 2, between Wave 2 and 3, and between Wave 1 and 3. More importantly, we computed test-retest correlations separately for females and males as well as for respondents of different age groups (defined by binning age at the time of the first data collection point into 10-year bins). Robustness checks (cf. Enkavi et al., 2019) suggested high correlations between test-retest correlations computed using different metrics and using (non)transformed data (Figures S2 and S3). Consequently, we report results using Pearson's r correlation coefficients for non-transformed data. To obtain reasonable estimates, test-retest correlations calculated from less than 30 responses were excluded from the main analyses. Further, we restricted the data set to correlations with a retest interval of up to 20 years. This resulted in a set of 72,963 test-retest correlations.

Variance Decomposition

To estimate the proportion of variance in the 72,963 test-retest correlations that could be explained by measure-related, respondent-related, and panel predictor variables, we used Shapley Decomposition (Grömping, 2007). First, we obtained the adjusted R^2 value from each of the 2^8 subsets of linear regression models (2^7 regression models for the category-specific variance decomposition). Second, we estimated the variance explained by each predictor by calculating the weighted average change in adjusted R^2 resulting from its inclusion in the model. Third, using 100 re-sampled data sets we generated 100 bootstrapped estimates for each prediction and from which we

computed bootstrapped confidence intervals (e.g., Sharapov et al., 2021).

Meta-Analytic Stability and Change Model

Model Description. The Meta-Analytic Stability and Change model (MASC) is a non-linear model introduced by Anusic and Schimmack (2016) to capture the trajectory of test-retest correlations over time. In this model, the test-retest correlation r_{t2-t1} at a specific *time* interval is a function of the proportion of reliable between-person variance, *rel*, the proportion of this reliable variance explained by changing factors, *change*, and the stability of these changing factors over time (per year), *stabch*. This is formalised as

$$r_{t2-t1} = rel \times (change \times (stabch^{\text{time}} - 1) + 1)$$

Figure S4A describes the model, and Figure S4B illustrates how different model parameterisations alter the shape of the curve.

Aggregation of Test-Retest Correlations. To minimise potential convergence issues that arise from meta-analysing 72,963 test-retest correlations using MASC, we aggregated the test-retest correlations. We obtained these aggregates by first grouping the test-retest correlations by sample, measure category, domain, and retest interval, as well as respondent gender and age group. We then calculated the average test-retest correlation for each of these groupings, using inverse-variance weighting and accounting for the dependency between these correlations. This resulted in 7,996 aggregated correlations.

Bayesian Model Specification. We set up the MASC model such that for each parameter (i.e., *rel*, *change* and *stabch*) we accounted for the effects of domain, linear age, quadratic age and gender, as well as the interaction between linear and quadratic age with domain. In addition, we included *sample* as a random factor for the *rel* parameter. Importantly, to obtain meta-analytic estimates we additionally specified the (aggregate) standard errors of each correlation. Lastly, to best capture domain-specific effects within each category, we fitted the model separately for each

measure category using their respective aggregated retest correlations and aggregated standard errors.

To estimate the parameters of this non-linear hierarchical model we used a Bayesian approach to account for the large differences between sample sizes and retest intervals encountered in such a large set of data sources. We specified weakly informative priors on the model parameters and hierarchical standard deviations so as to include values reported previously in the literature (e.g., Anusic and Schimmack, 2016; Frey et al., 2017; Mata et al., 2018).

Analyses were conducted in the R statistical environment (R Core Team, 2021), using the *brms* package (Bürkner, 2017, 2018, 2021) which provides a high-level interface to fit hierarchical models in Stan (Carpenter et al., 2017).

Construct Comparison. To compare the temporal stability and reliability of risk preference to that of other psychological constructs (e.g., personality), we re-analysed the set of correlations included in Anusic and Schimmack (2016) using a Bayesian estimation procedure and set of MASC model specifications to maximise comparability to the analyses conducted for risk preference.

Convergent Validity

In what follows, we give an overview of the main steps involved in computing inter-correlations between measures, variance decomposition of inter-correlations, and the meta-analyses of convergent validity. We provide additional information concerning each step in the Supplementary Information.

Computing Correlations

For the assessment of the convergence of risk preference measures, we started with the set of samples used to assess the temporal stability of risk preference, but selected only those samples that included two or more measures of risk preference within at least one wave, and for which the same set of respondents had provided answers. As a result, we conducted our convergent validity analyses for 45 samples from 26 panels (Figure 1), retaining the same three measure categories and 14 domains used

in the temporal stability analyses. First, for each sample, we computed the correlations between every possible pair of measures within the same data collection point. We computed these correlations separately for females and males as well as respondents of different ages. We excluded inter-correlations computed from the responses of less than 30 respondents. This resulted in a data set of 61,644 inter-correlations. Robustness checks (cf. Enkavi et al., 2019) suggested high correlations between inter-correlations computed using different metrics and using (non)transformed data (Figures S5 and S6). Here, we report results using Spearman's rho correlation coefficients for non-transformed data and which were based on a minimum of 30 responses.

To avoid model convergence issues when running the meta-analysis, we grouped the inter-correlations (e.g., by type of pair, age, gender, panel), and then aggregated the inter-correlations within these groupings, resulting in 5,038 aggregated inter-correlations.

Variance Decomposition

To estimate the proportion of variance in inter-correlations between risk preference measures that could be explained by measure-related, respondent-related, and panel predictor variables, we used Shapley Decomposition (Grömping, 2007). We followed the same approach used for the test-retest correlations obtaining the adjusted R^2 value from each of the (2^8) models, estimating the variance explained by each predictor by calculating the weighted average change in adjusted R^2 resulting from its inclusion in the model, and using a bootstrapping procedure to compute confidence intervals.

Meta-Analysis

To obtain the overall meta-analytic estimate of the convergence of risk preference measures, we first fitted a Bayesian hierarchical intercept-only model. Second, to obtain meta-analytic estimates for the convergence between specific pairs of measure categories and domains, we fitted Bayesian hierarchical (robust) regression models that included a predictor coding for the different types of measure pairs.

Multiverse Analyses

We conducted a series of multiverse analyses with alternative data sets resulting from different data pre-processing and various alternative analytic choices. We find overall qualitatively similar patterns of results across the multiverse of choices considered. We provide additional details concerning these analyses and results in the Supplementary Information.

Data and Code Availability

All the data are made publicly available through the original data repositories and need to be accessed by following the providers' data access policies. We provide more detailed overview of data, analysis, and code in a companion website (https://cdsbasel.github.io/temprisk/) and make the estimated test-retest correlations and inter-correlations from the primary data sources as well as all analysis scripts publicly available in an online repository (https://osf.io/5kzgd/).

Acknowledgements

This work was supported by grants from the Swiss National Science Foundation to R.M. (https://data.snf.ch/grants/grant/204700, https://data.snf.ch/grants/grant/177277). The authors thank Laura Wiles for editing the manuscript.

Author contributions

A.B.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing - original draft, and Writing - review & editing.

Y.L.: Data curation.

M.K.: Data curation.

G.S.: Data curation.

P.-C.B.: Formal analysis, Methodology, and Writing - review & editing.

L.T.: Conceptualization, Visualization, Writing - original draft, and Writing - review & editing.

R.M.: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Visualization, Writing - original draft, and Writing - review & editing.

Competing interests

The authors declare no competing interests.

References

Ahun, M. N., Lauzon, B., Sylvestre, M.-P., Bergeron-Caron, C., Eltonsy, S., & O'Loughlin, J. (2020). A systematic review of cigarette smoking trajectories in adolescents. *International Journal of Drug Policy*, *83*, 102838. https://doi.org/10.1016/j.drugpo.2020.102838

Anusic, I., & Schimmack, U. (2016). Stability and change of personality traits, self-esteem, and well-being: Introducing the meta-analytic stability and change model of retest correlations. *Journal of Personality and Social Psychology*, *110*(5), 766–781. https://doi.org/10.1037/pspp0000066

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. https://doi.org/10.1038/s41598-020-72077-5

Barseghyan, L., Molinari, F., O'Donoghue, T., & Teitelbaum, J. C. (2018). Estimating risk preferences in the field. *Journal of Economic Literature*, *56*(2), 501–564. https://doi.org/10.1257/jel.20161148

Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, *122*(3), 523–553. https://doi.org/10.1037/pspp0000386

Bernoulli, D. (1954). Exposition of a New Theory on the Measurement of Risk. *Econometrica*, *22*(1), 23–36. https://doi.org/10.2307/1909829

Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. A. (2022). Personality stability and change: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *148*, 588–619. https://doi.org/10.32614/RJ-2018-017

Bran, A., & Vaidis, D. C. (2020). Assessing risk-taking: What to measure and how to measure it. *Journal of Risk Research*, *23*(4), 490–503. https://doi.org/10.1080/13669877.2019.1591489

Breivik, G., Sand, T. S., & Sookermany, A. M. (2019). Risk-taking and sensation seeking in military contexts: A literature review. *SAGE Open*, *9*(1), 2158244018824498. https://doi.org/10.1177/2158244018824498

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to Basics: The Importance of Conceptual Clarification in Psychological Science. *Current Directions in Psychological Science*, *31*(4), 340–346. https://doi.org/10.1177/09637214221096485

Brodbeck, J., Duerrenberger, S., & Znoj, H. (2009). Prevalence rates of at risk, problematic and pathological gambling in Switzerland. *The European Journal of Psychiatry*, *23*(2), 67–75.

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

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411.

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*, 1–54. https://doi.org/10.18637/jss.v100.i05

Caliendo, M., Fossen, F., & Kritikos, A. S. (2014). Personality characteristics and the decisions to become and stay self-employed. *Small Business Economics*, *42*(4), 787–814. https://doi.org/10.1007/s11187-013-9514-8

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32. https://doi.org/10.18637/jss.v076.i01

Chapman, J., Dean, M., Ortoleva, P., Snowberg, E., & Camerer, C. (2018). *Econographics* (tech. rep. w24931). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w24931

Chuang, Y., & Schechter, L. (2015). Stability of experimental and survey measures of risk, time, and social preferences: A review and some new results. *Journal of*

Development Economics, *117*, 151–170.
https://doi.org/10.1016/j.jdeveco.2015.07.008

Conley, J. J. (1984). The hierarchy of consistency: A review and model of longitudinal findings on adult individual differences in intelligence, personality and self-opinion. *Personality and Individual Differences*, *5*(1), 11–25.
https://doi.org/10.1016/0191-8869(84)90133-8

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences*, *24*(4), 267–269.
https://doi.org/10.1016/j.tics.2020.01.007

Das, A., & Ahmed, M. M. (2022). Structural equation modeling approach for investigating drivers' risky behavior in clear and adverse weather using SHRP2 naturalistic driving data. *Journal of Transportation Safety & Security*.
https://doi.org/10.1080/19439962.2022.2155744

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

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259–268.
https://doi.org/10.1016/j.jrp.2011.02.004

Eisenberg, I. W., Bissett, P. G., Enkavi, A. Z., 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*(1), 2319.
https://doi.org/10.1038/s41467-019-10301-1

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychological Science*, *31*(7), 792–806.
https://doi.org/10.1177/0956797620916786

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477. https://doi.org/10.1073/pnas.1818430116

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, *133*(4), 1645–1692. https://doi.org/10.1093/qje/qjy013

Financial Services Authority. (2011). *Assessing suitability: Establishing the risk a customer is willing and able to take and making a suitable investment selection* (tech. rep.). Financial Services Authority.

Fraley, R. C., & Roberts, B. W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review*, *112*(1), 60–74. https://doi.org/10.1037/0033-295X.112.1.60

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, *3*(10), e1701381. https://doi.org/10/gb2xrw

Graham, E. K., Weston, S. J., Gerstorf, D., Yoneda, T. B., Booth, T., Beam, C. R., Petkus, A. J., Drewelies, J., Hall, A. N., Bastarache, E. D., Estabrook, R., Katz, M. J., Turiano, N. A., Lindenberger, U., Smith, J., Wagner, G. G., Pedersen, N. L., Allemand, M., Spiro, A., … Mroczek, D. K. (2020). Trajectories of Big Five personality traits: A coordinated analysis of 16 longitudinal samples. *European Journal of Personality*, *34*(3), 301–321. https://doi.org/10.1002/per.2259

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, *61*(2), 139–147. https://doi.org/10.1198/000313007X188252

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. (2020). Learning from the reliability paradox: How

theoretically informed generative models can advance the social, behavioral, and brain sciences. https://doi.org/https://doi.org.10.31234/osf.io/xr7y3

Hall, B. D., Liu, Y., Jansen, Y., Dragicevic, P., Chevalier, F., & Kay, M. (2022). A survey of tasks and visualizations in multiverse analysis reports. *Computer Graphics Forum*, *41*(1), 402–426. https://doi.org/10.1111/cgf.14443

Harrison, G. W. (2014). Real choices and hypothetical choices. In S. Hess & A. Daly (Eds.), *Handbook of choice modelling* (pp. 236–254). Edward Elgar Publishing.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis* (1st Ed.). Academic Press.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *The American Economic Review*, *92*, 1644–1655. https://doi.org/10.1257/000282802762024700

Jin, H., Cui, M., & Liu, J. (2020). Factors affecting people's attitude toward participation in medical research: A systematic review. *Current Medical Research and Opinion*, *36*(7), 1137–1143. https://doi.org/10.1080/03007995.2020.1760807

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*(3), 430–450. https://doi.org/10.1037/pspp0000090

Karlsson Linnér, R., Biroli, P., Kong, E., Meddens, S. F. W., Wedow, R., Fontana, M. A., Lebreton, M., Tino, S. P., Abdellaoui, A., Hammerschlag, A. R., Nivard, M. G., Okbay, A., Rietveld, C. A., Timshel, P. N., Trzaskowski, M., de Vlaming, R., Zünd, C. L., Bao, Y., Buzdugan, L., ... Beauchamp, J. P. (2019). Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics*, *51*(2), 245–257. https://doi.org/10.1038/s41588-018-0309-3

Karlsson Linnér, R., Mallard, T. T., Barr, P. B., Sanchez-Roige, S., Madole, J. W., Driver, M. N., Poore, H. E., de Vlaming, R., Grotzinger, A. D., Tielbeek, J. J., Johnson, E. C., Liu, M., Rosenthal, S. B., Ideker, T., Zhou, H., Kember, R. L., Pasman, J. A., Verweij, K. J. H., Liu, D. J., ... Dick, D. M. (2021). Multivariate

analysis of 1.5 million people identifies genetic associations with traits related to self-regulation and addiction. *Nature Neuroscience*, *24*(10), 1367–1376. https://doi.org/10.1038/s41593-021-00908-3

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

Levy, I., Snell, J., Nelson, A. J., Rustichini, A., & Glimcher, P. W. (2010). Neural Representation of Subjective Value Under Risk and Ambiguity. *Journal of Neurophysiology*, *103*(2), 1036–1047. https://doi.org/10.1152/jn.00853.2009

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. https://doi.org/10.1257/jep.32.2.155

Moffitt, T. E. (2018). Male antisocial behavior in adolescence and beyond. *Nature Human Behaviour*, *2*, 177–186. https://doi.org/10.1038/s41562-018-0309-4

Möttus, R., Briley, D. A., Zheng, A., Mann, F. D., Engelhardt, L. E., Tackett, J. L., Harden, K. P., & Tucker-Drob, E. M. (2019). Kids becoming less alike: A behavioral genetic analysis of developmental increases in personality variance from childhood to adolescence. *Journal of Personality and Social Psychology*, *117*(3), 635–658. https://doi.org/10.1037/pspp0000194

Norris, E., Finnerty, A. N., Hastings, J., Stokes, G., & Michie, S. (2019). A scoping review of ontologies related to human behaviour change. *Nature Human Behaviour*, *3*(2), 164–172. https://doi.org/10.1038/s41562-018-0511-4

Orth, U. (2018). Development of self-esteem from age 4 to 94 years: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *144*(10), 1045. https://doi.org/10.1037/bul0000161

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, *1*(11), 803–809. https://doi.org/10.1038/s41562-017-0219-x

R Core Team. (2021). R: A language and environment for statistical computing.

Ravert, R. D., Murphy, L. M., & Donnellan, M. B. (2019). Valuing risk: Endorsed risk activities and motives across adulthood. *Journal of Adult Development*, *26*(1), 11–21. https://doi.org/10/gmndcz

Richmond-Rakerd, L. S., D'Souza, S., Andersen, S. H., Hogan, S., Houts, R. M., Poulton, R., Ramrakha, S., Caspi, A., Milne, B. J., & Moffitt. (2020). Clustering of health, crime and social-welfare inequality in 4 million citizens from two nations. *Nature Human Behaviour*, *4*(3), 255–264. https://doi.org/doi:10.1038/s41562-019-0810-4

Roberts, D. K., Alderson, R. M., Betancourt, J. L., & Bullard, C. C. (2021). Attention-deficit/hyperactivity disorder and risk-taking: A three-level meta-analytic review of behavioral, self-report, and virtual reality metrics. *Clinical Psychology Review*, *97*, 102039. https://doi.org/10.1016/j.cpr.2021.102039

Schildberg-Hörisch, H. (2018). Are risk preferences stable? *Journal of Economic Perspectives*, *32*(2), 135–154. https://doi.org/10.1257/jep.32.2.135

Schmidt, L. (2008). Risk preferences and the timing of marriage and childbearing. *Demography*, *45*(2), 439–460. https://doi.org/10.1353/dem.0.0005

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

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Seifert, I. S., Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2022). The development of the rank-order stability of the Big Five across the life span. *Journal of Personality and Social Psychology*, *122*, 920–941. https://doi.org/10.1037/pspp0000398

Sharapov, D., Kattuman, P., Rodriguez, D., & Velazquez, F. J. (2021). Using the SHAPLEY value approach to variance decomposition in strategy research: Diversification, internationalization, and corporate group effects on affiliate profitability. *Strategic Management Journal*, *42*(3), 608–623. https://doi.org/10.1002/smj.3236

Stan Development Team. (2022). Stan user's guide. Version 2.29.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Steinberg, L. (2013). The influence of neuroscience on US Supreme Court decisions about adolescents' criminal culpability. *Nature Reviews Neuroscience*, *14*(7), 513–518. https://doi.org/10/gdcf6b

Steiner, M. D., Seitz, F. I., & Frey, R. (2021). Through the window of my mind: Mapping information integration and the cognitive representations underlying self-reported risk preference. *Decision*, *8*, 97–122. https://doi.org/10.1037/dec0000127

Stigler, G. J., & Becker, G. S. (1977). De gustibus non est disputandum. *The American Economic Review*, *67*(2), 76–90. https://doi.org/http://www.jstor.org/stable/1807222

Strickland, J. C., & Johnson, M. W. (2021). Rejecting impulsivity as a psychological construct: A theoretical, empirical, and sociocultural argument. *Psychological review*, *128*(2), 336–361. https://doi.org/10.1037/rev0000263

Tisdall, L., MacNiven, K. H., Padula, C. B., Leong, J. K., & Knutson, B. (2022). Brain tract structure predicts relapse to stimulant drug use. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(26), e2116703119. https://doi.org/10.1073/pnas.2116703119

Williams, D. R., Rast, P., & Bürkner, P.-C. (2018). Bayesian meta-analysis with weakly informative prior distributions. https://doi.org/10.31234/osf.io/7tbrm

Wulff, D. U., & Mata, R. (2022). On the semantic representation of risk. *Science Advances*, *8*(27), eabm1883. https://doi.org/10.1126/sciadv.abm1883

Table 1

Descriptions and examples of different categories of risk preference measures

Category	Description	Example
Propensity	self-report measures; individuals indicate on a (ordinal) scale to what extent they identify as someone who likes or is willing to take risks in general or in specific domains.	*Are you generally a person who is willing to take risks or do you try to avoid taking risks?* (Dohmen et al., 2011)
Frequency	self-report measures; individuals indicate on a scale or in an open field to what extent or how often they partake in activities in specific life domains.	*How many times in the last seven days have you had an alcoholic drink?*, (Frey et al., 2017)
Behavioural	behavioural measures; individuals are asked to decide between two or more options typically offering different (hypothetical or real) monetary gains and/or losses with varying probability; an index of risk preference is typically derived based on a combination of choices or actions.	Mean number of pumps in a simulated balloon-pumping task (Lejuez et al., 2002); percentage of risky choices in a lottery task (Holt & Laury, 2002)

Figure 1

Flowchart of systematic search.

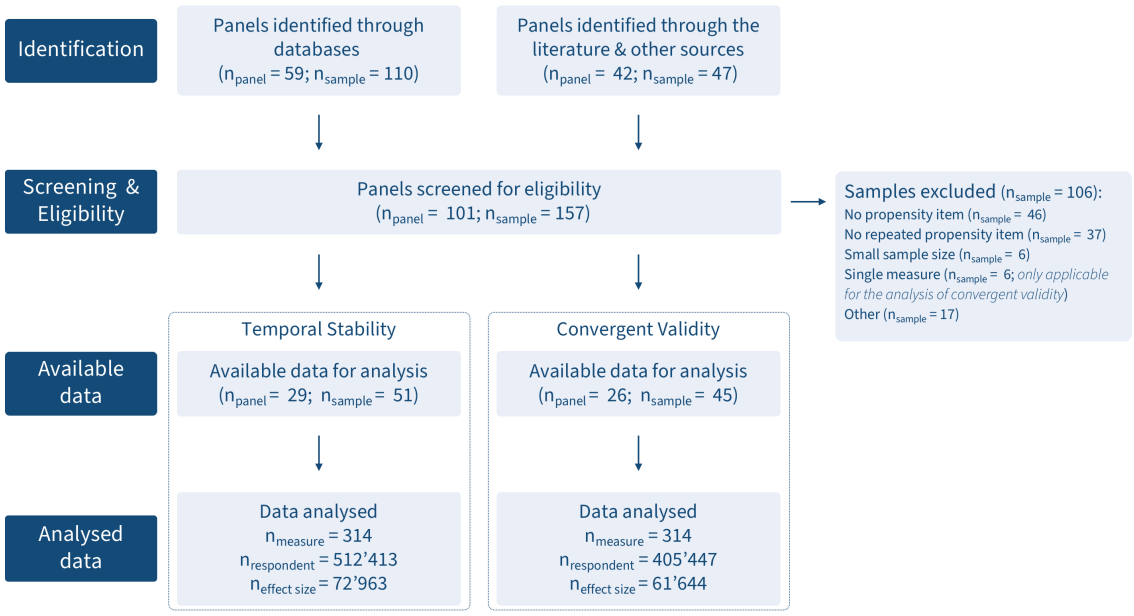


Figure 2

Overview of data. A) Two Dimensional density plot of test-retest correlations as a function of retest interval (k = 72,963). B) Distribution of all inter-correlations (k = 61,644). C) Number of unique measures split by category (propensity, frequency, behaviour), and domain.

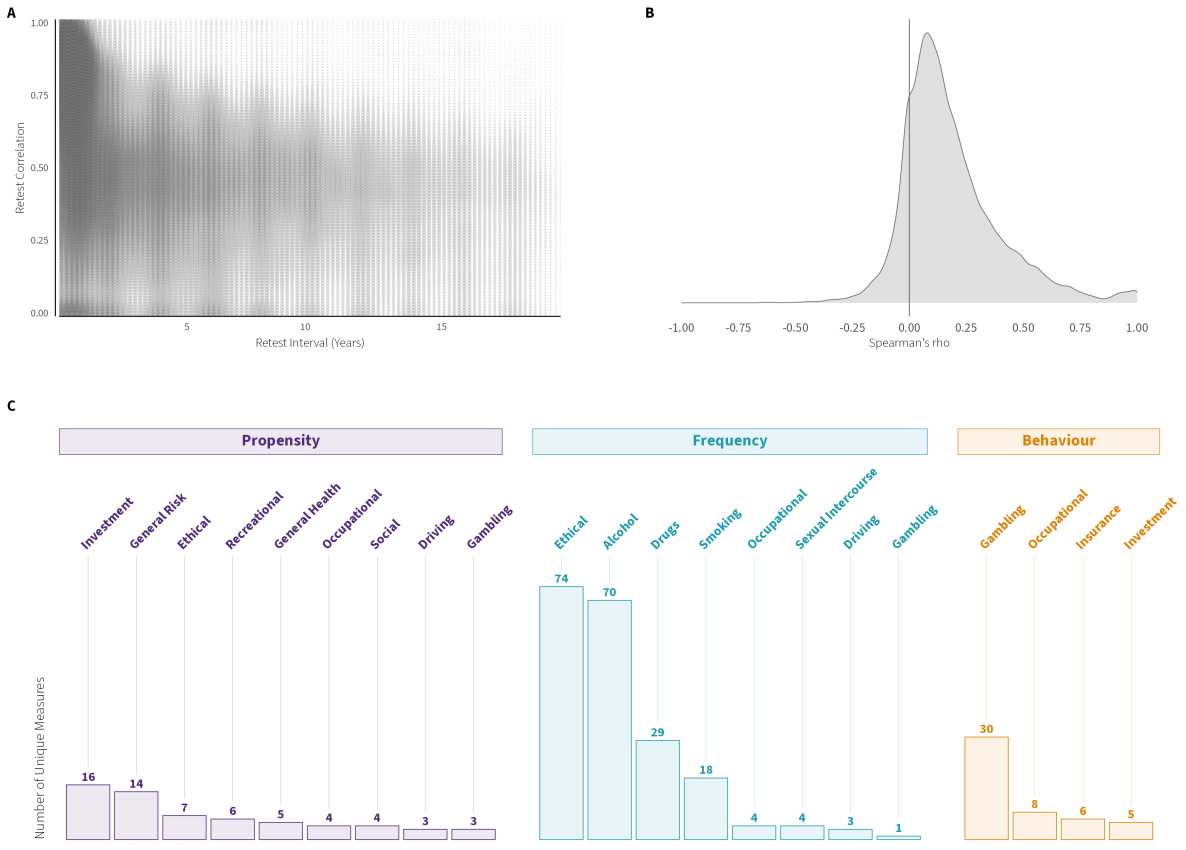


Figure 3

Variance decomposition of temporal stability. A) Relative contribution of measure, respondent, and panel-related predictors to the adjusted R^2 in regression models predicting test-retest correlations of all risk preference measures (k = 72,963). B) Relative contribution of measure, respondent, and panel predictors to the adjusted R^2 in regression models predicting test-retest correlations of propensity (k = 23,936), frequency (k = 47,490), and behavioural (k = 1,537) measures. Estimate (dot) and bootstrapped (coloured area) 95%, 80%, and 50% confidence intervals.

Figure 4

Meta-Analytic Stability and Change model (MASC) results. The figure shows parameter estimates for A) propensity (k = 3,706), B) frequency (k = 3,678), and C) behavioural measures (k = 612) of risk preference. In A-C, circles represent the mean estimate, shaded uncertainty bands represent the 50%, 80%, and 95% HDI. D-I show predictions of retest trajectories given MASC parameters as a function of retest interval (D,F,H) or age (E,G,I) across all domains (shaded uncertainty bands, 50%, 80%, and 95% HDI) as well as a selection of two domains per category (individual, annotated lines)

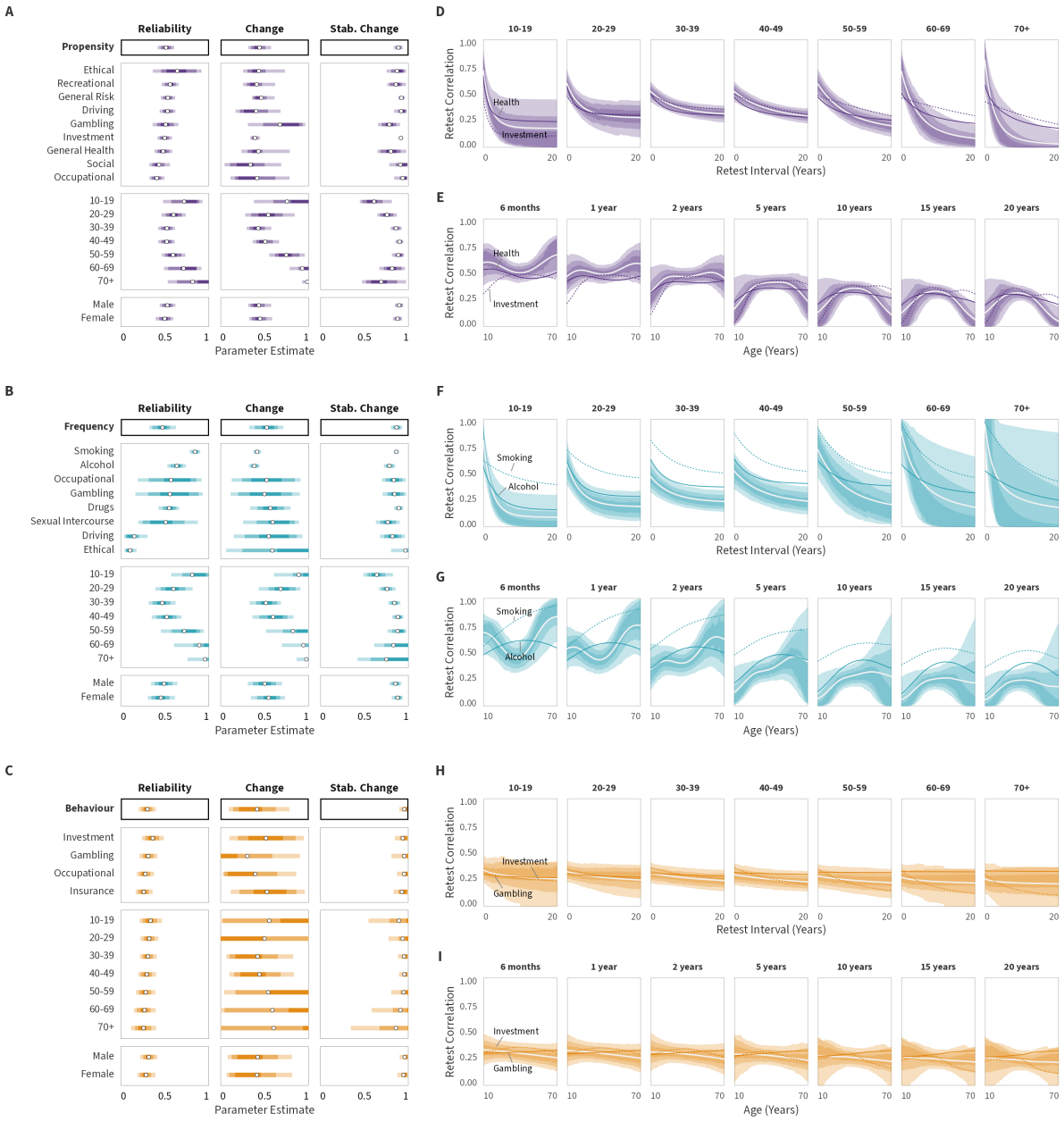


Figure 5

Variance decomposition of convergence between measures. Relative contribution of measure, respondent, and panel-related predictors to the adjusted R^2 in regression models predicting inter-correlations between measures of risk preference (k = 61,644). Estimate (dot) and bootstrapped (coloured area) 95%, 80%, and 50% confidence intervals.

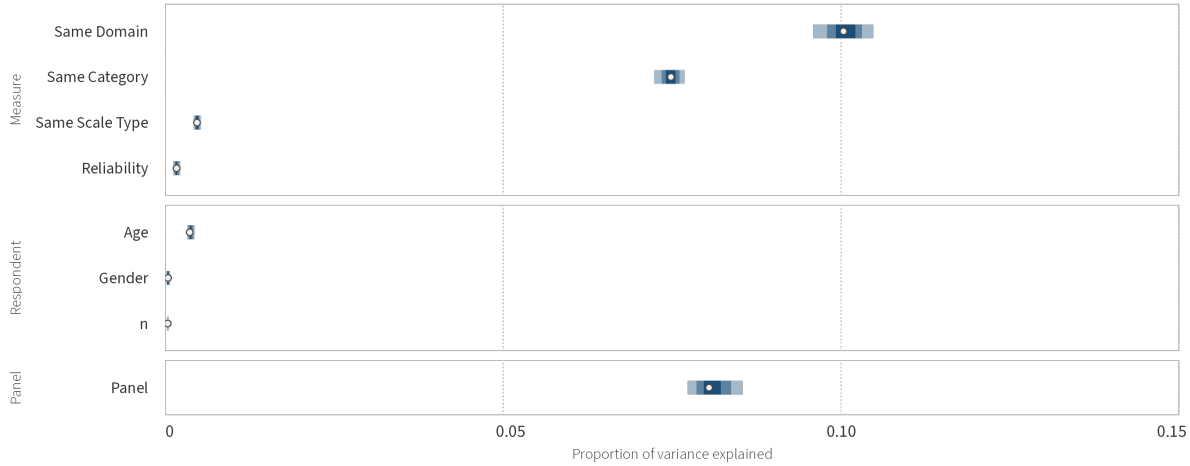


Figure 6

Meta-analytic correlation matrices. The matrices depicts the results of the meta-analyses of inter-correlations between measures of risk preference (k = 5,038), with each cell representing the meta-analytic result for the specific measurement pair of A) measure domains or B) measure categories. Empty grey cells are due to lack of data availability to estimate the respective correlation.

Supplementary Information

*

Contents

Identification of Samples S2
Categorisation of Measures S2
Data Pre-Processing S3
Data Set Information S3
Risk Preference Measures S3
Variable recoding based on question dependency. S3
Reverse coding. S4
Composite measures. S4
Harmonising variable names S4
Sample Demographics S4
Data Processing S5
Temporal Stability S5
Computing test-retest correlations S5
Aggregating test-retest correlations S7
Convergent Validity S7
Computing inter-correlations S7
Aggregating inter-correlations S8
Analysis S9
Temporal Stability S9
Variance decomposition S9
Meta-Analytic Stability and Change model (MASC) S10
Re-analysis of the Anusic and Schimmack (2016) data set S13
Convergent Validity S14
Variance decomposition S14
Meta-analyses S14
Multiverse Analyses S16

Identification of Samples

To find as many longitudinal panels and associated samples with risk preference measures, we devised a list of search terms related to risk (e.g., risk*, gambl*, smok*, gambl*; Table S1). This list reflects the definition of risk from the economics and psychology literatures and covers many different areas of life. It was developed by consulting the questionnaires of multi-measure studies (Arslan et al., 2020; Chapman et al., 2018; Eisenberg et al., 2019; Enkavi et al., 2019; Falk et al., 2018; Frey et al., 2017) as well as previously identified longitudinal samples (e.g., SOEP, USOC). As presented in the main paper, this search led us to identify a large number of panels (101) and associated samples (157) (Table S2), which we checked for possible inclusion in our study. We excluded sample or measures from our study using a clear set of inclusion/exclusion criteria (Table S3). Each sample was documented differently, thus, whenever available, we used the computerised (online) variable search engine to search for the risk-related terms, otherwise, we manually searched the codebooks and/or questionnaires available. Our systematic approach to search and screening resulted in the inclusion of 51 unique samples from 29 panels (Table S4).

Categorisation of Measures

We conducted extensive coding and categorisation of each risk preference measure that met our inclusion criteria. Specifically, we coded the following information: the name of the panel it originated from, the measure category (i.e., propensity, frequency, or behaviour), the domain (e.g., recreational, smoking), the type of scale used (i.e., ordinal, discrete, composite or open ended) and, if ordinal or discrete (with a clear range of possible response values), the number of options or points in the scale. In addition, we included information that was specific to each category of risk preference measure. Specifically, for frequency measures, we specified the number of days over which a certain behaviour had to be reported (e.g., Over the last week/month/year how many times were you intoxicated?). For behavioural measures, we recorded whether the decision was incentivised or hypothetical (cf., Harrison, 2014). Please note that we do

not include these category-specific characteristics in our analyses because they are not instrumental to the comparison between categories. Nevertheless, we provide this categorisation for completeness and future possible uses of these data that control for or examine the role of such characteristics. Overall, we identified 314 unique measures stemming from 51 longitudinal samples. We provide a detailed definition, coding and description of each type of measure in Table S5, as well as a complete list of the risk preference measures in the main code book available in the online repository.

Data Pre-Processing

Prior to computing test-retest correlations, we pre-processed the data from each sample to create homogeneous data sets with regards to the data set information, risk preference measures, and sample demographics. We provide details concerning each step below.

Data Set Information

From each data set, we extracted the wave identifiers and data collection dates (i.e., day-month-year). If these dates were missing, we determined for each wave a standard date by referring to the sample's data collection timeline and choosing the half-way point (e.g., if data collection took place between January and June of 2020, the 15th of March 2020 was selected as the date). In the case that only the year could be retrieved, we set June 15th as the default day. If the data collection date was missing for certain respondents within the wave of a panel, this date was filled by the mean of the available dates.

Risk Preference Measures

Variable recoding based on question dependency. Depending on the design of the questionnaires/interviews, for some samples, respondents were not asked certain questions because of their response to previous (filter) questions. This was particularly the case for frequency measures. For instance, if an individual answered the question *"Are you currently a smoker?"* with *"No"*, the follow-up question *"How many*

cigarettes a day do you smoke?" would not be asked and would automatically receive a "missing" or "not applicable" code. By ignoring dependencies between questions, valuable information on the consistency of an individual's behaviour is missed, as instances of when behaviours might be interrupted and taken up again (e.g., quitting/taking up smoking) are unaccounted for. To deal with this, for each sample, we took into account responses to filter-type questions and replaced invalid/missing codes in subsequent related questions by an appropriate response. In the case of the above example, for all the participants who answered *"No"*, we replaced the invalid or missing code for the number of cigarettes smoked in a day with a *"0"* or *"None"*. To make such replacements possible, we only included measures in our analyses that had scales that offered the possibility of a 0 value or Never/None answer (Table S3).

Reverse coding. Whenever appropriate, we reversed the scales of measures such that higher values corresponded to greater risk-taking.

Composite measures. We define a composite measure as a measure which represents an index of risk taking that is calculated by combining two or more individual risk preference measures. This was particularly the case for behavioural measures. If a composite measure was not available in the raw data set of the sample, we aggregated the set of available single responses using similar methods as that of studies with comparable tasks (e.g., proportion of risky choices). We provide a description of how these have been calculated for specific measures in the risk preference measure code book.

Harmonising variable names. We standardised the names of the measures such that the same risk preference measure (or highly similarly worded measure with the same response format and scale) included in different samples shared the same variable name.

Sample Demographics

We recorded the age and gender of each respondent. Age was calculated at the time of each data collection point. If the respondent's birth year was available in the data set, we used that to calculate their age, if not, we used the value of the

pre-computed age in the data set. Further, if only age group or age range information was available (e.g., 20-30), we defined age as the midpoint value (e.g., 25). Only data from respondents between the ages of 10 and 90 years were included in the analyses. We coded gender as a binary variable (0 = male and 1 = female). For data quality purposes, we did not include in our analyses the responses of respondents whose year of birth, age (i.e., if the age difference and time difference between first and last wave of participation differed by more than 2 years) or gender was inconsistently reported across waves. Additionally, if either the year of birth, age or gender was missing and could not be retrieved or estimated based on previous waves, the respondent was excluded from the analyses.

Data Processing

Temporal Stability

Computing test-retest correlations. To address our main research objectives, for each panel and risk preference measure, we calculated for all possible wave combinations test-retest correlations (Figure S1). Correlations were calculated separately for females and males of different age groups. We computed separate sets of test-retest correlations for different age group configurations: 5, 10 and 20-year age bins. Akin to Enkavi et al. (2019), we estimated test-retest correlations using three different metrics: Pearson's r, Spearman's rho and intra-class correlations (ICC(2,1)). The correlation between these different metrics ranged between 0.58 and 0.99 (Figure S2). Further, the response distributions of some measures were highly skewed, thus we additionally computed test-retest correlations using log-transformed data. As shown in Figure S3, these were highly correlated with the test-retest correlations computed using the non-transformed data (r = 0.91 - 0.98). As a consequence, we report our main results using the Pearson's r correlation coefficient for the non-transformed data. Furthermore, when computing the test-retest correlations we obtained negative estimates (3.95% of the data set used for analysis); for ease of interpretation, we replaced these values with zeroes prior to any analysis or aggregation procedures (cf.,

Enkavi et al., 2019).

Additional metrics. In addition to these correlation metrics, for each test-retest correlation coefficient we recorded the following (variables with an asterisk were included in our main analyses, the rest were included for data quality assessment and data exploration):

- Respondent information: sample size, maximum age, minimum age, mean age*, median age, standard deviation of age, proportion of female respondents*, proportion of sample lost between the first and second data collection point (i.e., attrition rate)
- Retest interval: minimum, maximum, mean*, median and standard deviation of the number of years between the first and second data collection point
- Response properties: the coefficient of variation and skewness of the responses at both time points

When calculating the time interval between the first and second data collection point, we noted that for panels that collected data for different surveys simultaneously (e.g. American Life Panel), not all respondents completed the surveys in the same order; some respondents would complete a more recent survey prior to an older survey (based on the mean data collection date), resulting in a negative retest interval. Therefore, for a very small number of correlations (0.17%) the minimum retest interval was negative. However, in our analyses we use the mean time difference between waves (or surveys), which minimises this issue. One exception to this concerns the German Socioeconomic Panel, which in 2020 launched a COVID-specific survey in which data collection overlapped with the 2020 core survey. We could not adequately order this pair of waves (i.e., 2020-core and 2020-covid) as we systematically had correlations that either had a negative mean or median retest interval. Therefore, we excluded correlations that from this specific pair of waves.

Sample size. Simulation studies have shown that large sample sizes may be needed to compute stable correlation coefficients (Schönbrodt & Perugini, 2013). On the

companion website we show how the number of correlation coefficients in the data set varies for different age groups based on different minimum sample size thresholds. For some age groups a substantial number of coefficients are lost as the threshold increases. To avoid losing valuable information for certain age groups, we retained the set of test-retest correlations that had a sample size of at least 30 with age groups organised in 10-year bins. In line with the multiverse approach (Steegen et al., 2016), the companion website provides an overview of the outcome of our analysis obtained using the different minimum sample size thresholds, age bins, and other processing steps.

Aggregating test-retest correlations. Given the high number of test-retest correlations in our data set (N = 72,963 correlations), it was too complex and computationally intensive to adequately estimate the Meta-Analytic Stability and Change model (MASC; Anusic and Schimmack (2016)) and capture the trajectories of the correlations over time without encountering severe model convergence issues. Therefore, we aggregated the correlations prior to fitting the MASC model. Specifically, first, we transformed each Pearson's r correlation coefficient into Fisher's z, and calculated the corresponding sampling variance. Second, we grouped the test-retest correlations by panel, measure category, measure domain, 3-month retest interval, gender, and age group. For each grouping we computed a synthesised estimate by aggregating test-retest correlation coefficients whilst accounting for the dependency between them as these were computed from the same set or subset of respondents (Hedges & Olkin, 1985). For this purpose, we used inverse-variance weighting and set the correlation of the sampling errors within subsets to .5. Lastly, these aggregated correlations and their standard errors were back transformed to Pearson's r. Given that MASC model predictions are bounded between 0 and 1, we set any negative aggregated retest correlation to zero. This process resulted in 7,996 aggregated test-retest correlation being calculated.

Convergent Validity

Computing inter-correlations. Samples which contained only one measure of risk preference were excluded from these analyses (n = 6). For each of the remaining

samples and waves, we calculated correlations between the responses of every possible pair of measures, for every wave same time point. Similar to the test-retest correlations, inter-correlations were calculated separately for females and males of different age groups. Specifically, we computed separate sets of correlations for different age group configurations: 5, 10 and 20-year age bins. We estimated inter-correlations using three different metrics, Pearson's r, Spearman's rho and intraclass correlations (ICC(2,1)), and examined inter-correlations being computed using non-transformed or log-transformed data. As shown in Figure S5, inter-correlations computed using different metrics were highly correlated (r = 0.84 - 0.92), as were the inter-correlations for (non)transformed data (r = 0.95 - 0.99) (Figure S6).

Additional metrics. For each inter-correlation coefficient we additionally recorded the following (variables with an asterisk were included in our main analyses, the rest were included for data quality assessment and data exploration):

- Response information: sample size, maximum age, minimum age, mean age*, median age, standard deviation of age, proportion of female respondents*
- Response properties: the coefficient of variation and skewness of the responses of both measures

Aggregating inter-correlations. In an effort to reduce computational costs and the potential occurrence of divergent transitions when conducting the Bayesian meta-analysis, we aggregated the inter-correlations. We followed a similar approach as for the retest correlations, we first converted each correlation coefficient into Fisher's z, and calculated the corresponding sampling variance. We then split the set of inter-correlations by sample, gender, age group, and category-domain measure pairs. For each subset we computed a synthesised estimate by aggregating the Fisher's z values using inverse-variance weighting and accounting for the dependency between them as these were computed from the same set or subset of respondents (Hedges & Olkin, 1985). To average these correlations we used inverse-variance weighting and set the correlation of the sampling errors within subsets to .5. We conducted additional

analyses in which we tested the effects of this correlation on our results by setting the correlation to 0.1 and 0.9.

Analysis

Temporal Stability

Variance decomposition. To gain a better understanding of the heterogeneity observed between test-retest correlations, we conducted a variance decomposition analysis by computing the Shapley values for the following predictors:

Measure characteristics

- Category: type of measure (i.e., propensity, frequency, behaviour)
- Domain: life domain the measure focuses on (e.g., smoking, driving, social, ethical)
- Scale type: type of response scale (i.e., open-ended/composite index, ordinal/discrete scales)
- Retest Interval: number of years between T1 and T2 data collection

Respondent characteristics

- Age: age group the respondents belong in (10 year bins, e.g., 20-29, 30-39)
- Gender: gender of the respondents (i.e., female, male)
- Number of responses: sample size for each correlation

Shapley values were computed by first estimating a linear regression for each possible combination of predictors (i.e., 2^8 models for the omnibus analysis, and 2^7 models for the category-specific analyses) and extracting the adjusted R^2 value. Then, for each predictor, we computed the weighted average of the change in adjusted R^2 resulting from the inclusion of that predictor in the models.

To obtain bootstrapped confidence intervals, we sampled the data set of correlations 100 times, and estimated for each predictor a set of 100 Shapley values. To

visualise these results, we ranked these values to determine the 50%, 80% and 95% confidence intervals.

Meta-Analytic Stability and Change model (MASC).

Model specification. To assess the trajectory of test-retest correlations of risk preference over time we used the MASC model developed by Anusic and Schimmack (2016) (Figure S4). Specifically, we were interested in quantifying the effects of gender, linear age, quadratic age, and domain, as well as the interactions between linear and quadratic age with domain on each of the MASC model parameters (i.e, *rel*, *change* and *stabch*).

In the model, domain was a sum contrast coded factor, gender was the proportion of female respondents (*FemaleProp*) centred at 0.5 (i.e., -0.5 = males and 0.5 = females), and age (*Age*) corresponded to the mean age of the respondents centred at 40 years and transformed into decades. Quadratic age (*Age2*) was the square value of the *Age* predictor. Lastly, retest interval was coded as the number of decades between waves.

The samples differed from each other on multiple dimensions (e.g, country, mode of data collection), hence, to account for such differences when estimating the MASC model parameters, we included *sample* as a random factor. We limited the (correlated) random effects structure to the *rel* parameter by adding a varying intercept and varying slopes for the effects of linear age, quadratic age and gender[1]. We did not include a random effects structure for the estimation of the *change* and *stabch* parameters, because to appropriately estimate these parameters samples should have data for a long enough period such that the test-retest correlations assymptote (Anusic & Schimmack, 2016). In the current data set, the number of test-retest correlations per sample varied substantially, and less than the majority of the samples ($\sim$ 40%) contained retest correlations beyond an interval of 10 years.

The values of *rel*, *change* and *stabch* are bounded between 0 and 1. The *rel* and

[1] We did not include a varying slope for the effect of domain as not every sample had data on each level of domain.

change parameters both represent proportions (i.e., the proportion of reliable between-person variance and the proportion of reliable variance attributable to changing factors, respectively). For the *stabch* parameter (i.e., the rate of change) we need to take into account that over the years changes in individuals' lives accumulate and gradually affect their behaviour to different extents, resulting in decreasing (i.e., 0 $<$ rate of change $\leq$ 1) rather than increasing (i.e., rate of change $>$ 1) correlations across the years (Anusic & Schimmack, 2016). Therefore, to ensure that these parameters remained within their valid intervals, we modelled them on the logit scale (i.e., *logitrel*, *logitchange* and *logitstabch*), and subsequently back-transformed them via the inverse logit function (Bürkner, 2021). Such as to obtain meta-analytic estimates of each parameter, we additionally specified in the model the corresponding standard errors of the (aggregated) retest correlations.

We used Bayesian inference to estimate the meta-analytic model and specified weakly informative priors for the model parameters and hierarchical standard deviations so as to include estimates reported in previous literature (e.g., Anusic and Schimmack, 2016; Frey et al., 2017; Mata et al., 2018). The Bayesian hierarchical non-linear model described below was estimated using the probabilistic programming language Stan (Carpenter et al., 2017; Stan Development Team, 2022) via the R package *brms* (Bürkner, 2017, 2018, 2021). The companion website reports the summary output of the model, sample-specific model predictions, and MCMC diagnostic plots.

$$
\begin{aligned}
y_i &\sim StudentT(\nu, \theta_i, \sqrt{se_i^2 + \sigma^2}) \\
\theta_i &= rel_i \times (change \times (stabch^{\text{time}} - 1) + 1) \\
rel_i &= logit^{-1}(logitrel_i) \\
change &= logit^{-1}(logitchange) \\
stabch &= logit^{-1}(logitstabch) \\
\sigma &\sim Cauchy_+(0, 1) \\
\nu &\sim Gamma(2, 0.1)
\end{aligned}
$$

logitrel$_i$ parameter

$$logitrel_i = (\beta_{logitrel_0} + \beta_{logitrel_{0,sample[i]}}) + (\beta_{logitrel_1} + \beta_{logitrel_{1,sample[i]}}))Age+$$
$$(\beta_{logitrel_2} + \beta_{logitrel_{2,sample[i]}})Age2 + (\beta_{logitrel_3} + \beta_{logitrel_{3,sample[i]}})FemaleProp+$$
$$\beta_{logitrel_4}Domain + \beta_{logitrel_5}(Age \times Domain) + \beta_{logitrel_6}(Age2 \times Domain)$$

$$\beta_{logitrel_0}, \quad \beta_{logitrel_1}, \quad \beta_{logitrel_2}, \quad \beta_{logitrel_3}, \quad \beta_{logitrel_4}, \quad \beta_{logitrel_5}, \quad \beta_{logitrel_6} \sim Normal(0,1)$$

$$\begin{bmatrix} \beta_{logitrel_{0,sample}} \\ \beta_{logitrel_{1,sample}} \\ \beta_{logitrel_{2,sample}} \\ \beta_{logitrel_{3,sample}} \end{bmatrix} \sim MVNormal\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad Cov \right)$$

$$Cov = \begin{pmatrix} \sigma^2_{\beta_{logitrel_{0,sample}}} & \sigma_{\beta_{logitrel_{0,sample}}}\sigma_{\beta_{logitrel_{1,sample}}}\rho_{0,1} & \sigma_{\beta_{logitrel_{0,sample}}}\sigma_{\beta_{logitrel_{2,sample}}}\rho_{0,2} & \cdots \\ \sigma_{\beta_{logitrel_{1,sample}}}\sigma_{\beta_{logitrel_{0,sample}}}\rho_{0,1} & \sigma^2_{\beta_{logitrel_{1,sample}}} & \sigma_{\beta_{logitrel_{1,sample}}}\sigma_{\beta_{logitrel_{2,sample}}}\rho_{1,2} & \\ \sigma_{\beta_{logitrel_{2,sample}}}\sigma_{\beta_{logitrel_{0,sample}}}\rho_{0,2} & \sigma_{\beta_{logitrel_{2,sample}}}\sigma_{\beta_{logitrel_{1,sample}}}\rho_{1,2} & \sigma^2_{\beta_{logitrel_{2,sample}}} & \\ \sigma_{\beta_{logitrel_{3,sample}}}\sigma_{\beta_{logitrel_{0,sample}}}\rho_{0,3} & \sigma_{\beta_{logitrel_{3,sample}}}\sigma_{\beta_{logitrel_{1,sample}}}\rho_{1,3} & \sigma_{\beta_{logitrel_{3,sample}}}\sigma_{\beta_{logitrel_{2,sample}}}\rho_{2,3} & \end{pmatrix}$$
$$\begin{matrix} \sigma_{\beta_{logitrel_{0,sample}}}\sigma_{\beta_{logitrel_{2,sample}}}\rho_{0,3} \\ \sigma_{\beta_{logitrel_{1,sample}}}\sigma_{\beta_{logitrel_{3,sample}}}\rho_{1,3} \\ \cdots \quad \sigma_{\beta_{logitrel_{2,sample}}}\sigma_{\beta_{logitrel_{3,sample}}}\rho_{2,3} \\ \sigma^2_{\beta_{logitrel_{3,sample}}} \end{matrix}$$

$$\sigma_{\beta_{logitrel_{0,sample}}}, \quad \sigma_{\beta_{logitrel_{1,sample}}}, \quad \sigma_{\beta_{logitrel_{2,sample}}}, \quad \sigma_{\beta_{logitrel_{3,sample}}} \sim Cauchy_+(0,1)$$

$$\rho \sim LKJCorr(1)$$

logitchange parameter

$$logitchange = \beta_{logitchange_0} + +\beta_{logitchange_1}Age + \beta_{logitchange_2}Age2 + \beta_{logitchange_3}FemaleProp+$$
$$\beta_{logitchange_4}Domain + \beta_{logitchange_5}(Age \times Domain) + \beta_{logitchange_6}(Age2 \times Domain)$$

$$\beta_{logitchange_0}, \beta_{logitchange_1}, \beta_{logitchange_2}, \beta_{logitchange_3},$$
$$\beta_{logitchange_4}, \beta_{logitchange_5}, \beta_{logitchange_6} \sim Normal(0,1)$$

logitstabch parameter

$$logitstabch = \beta_{logitstabch_0} + +\beta_{logitstabch_1}Age + \beta_{logitstabch_2}Age2 + \beta_{logitstabch_3}FemaleProp+$$
$$\beta_{logitstabch_4}Domain + \beta_{logitstabch_5}(Age \times Domain) + \beta_{logitstabch_6}(Age2 \times Domain)$$

$$\beta_{logitstabch_0}, \beta_{logitstabch_1}, \beta_{logitstabch_2}, \beta_{logitstabch_3},$$

$$\beta_{logitstabch_4}, \beta_{logitstabch_5}, \beta_{logitstabch_6} \sim Normal(0, 1)$$

Re-analysis of the Anusic and Schimmack (2016) data set. We re-analysed the data that the authors made available in the study's supplementary material. The authors collated and analysed test-retest correlations spanning 15 years for assessments of personality traits, self-esteem, life satisfaction, and affect. Prior to any data processing or analysis we excluded from the data set retest correlations that were computed from samples that had missing sample size information (n = 4), and where respondents were on average below 10 years of age or above 90 years of age (n = 31) leaving a total of 949 test-retest correlations (personality = 226, self-esteem = 196, affect = 101, life satisfaction = 426) for analysis. To remain consistent with how we analysed the other set of retest correlations, prior to estimating the model parameters, we first:

a) calculated the sampling variance of each correlation using the following formula,

$$\frac{\sqrt{(1 - \text{retest}^2)^2}}{n - 1} \tag{1}$$

b) centered the age variable at 40 years and transformed it into decades,

c) centered the proportion of females variable at 0.5, and

d) rounded the retest interval variable to .25 (i.e., 3 months bins).

Given that in the data set close to 80% of the studies/samples had 4 or less observations, to avoid poor estimation of varying intercepts and slopes as well as model convergence issues, we did not specify a random effects structure for the *rel* parameter.

By following these data processing and analysis steps we deviated from the original study's analysis in four ways. First, we used a smaller data set. Second, we carried out the analysis using a Bayesian instead of a Frequentist approach. Third, when conducting the meta-analysis we accounted for the correlations' standard error. Lastly, we changed the moderators that were included in the model by adding an interaction between age linear and construct, between age quadratic and construct, and

removing the effect of scale length on the *rel* parameter.

Details of the model specification in *brms*, model fit and convergence statistics are provided in the companion website.

Convergent Validity

Variance decomposition. To gain a better understanding of the heterogeneity in the correlation between different measures, we conducted a variance decomposition analysis. We computed the Shapley values of the following predictors:

Measure characteristics

- Measure category match: whether or not both measures belong to the same category (i.e., propensity, frequency, behaviour)
- Domain match: whether or not both measures focus on the same life domain (e.g., smoking, driving, social, ethical)
- Scale type match: whether or not both measures have the same type of scale (i.e., open-ended/composite index, ordinal/discrete scales)
- Reliability: the average reliability of the measures (using MASC model parameter estimates to make measure and age-specific predictions)

Respondent characteristics

- Age: age group the respondents belong in (10 year bins)
- Gender: gender of the respondents (i.e., female, male)
- Number of responses: sample size for each correlation

Meta-analyses. Using the aggregated Fisher's z-transformed correlations, we conducted a Bayesian random-effects meta-analysis to quantify the convergence across all measures, and followed a distributional modelling approach by allowing the samples to vary in their residual standard deviation (σ).

$$
\begin{aligned}
y_i &\sim StudentT(\nu, \theta_i, \sqrt{se_i^2 + \sigma_i^2}) \\
\theta_i &\sim Normal(\mu_\theta, \tau_\theta) \\
\mu_\theta &\sim Normal(0, 1) \\
\tau_\theta &\sim Cauchy_+(0, 0.3) \\
\log\sigma_i &\sim Normal(\mu_\sigma, \tau_\sigma) \\
\mu_\sigma &\sim Normal(0, 2) \\
\tau_\sigma &\sim Cauchy_+(0, 0.3) \\
\nu &\sim Gamma(2, 0.1)
\end{aligned}
$$

Second, we conducted two meta-regressions with categorical covariates to estimate the convergence between a) different pairs of measure categories (e.g., frequency and propensity),

$$
\begin{aligned}
y_i &\sim StudentT(\nu, \theta_i, \sqrt{se_i^2 + \sigma_i^2}) \\
\theta_i &= \beta_{\theta_{0,sample[i]}} + \beta_{\theta_1} CategoryPair \\
\beta_{\theta_1} &\sim Normal(0, 1) \\
\beta_{\theta_{0,sample}} &\sim Cauchy_+(0, 0.3) \\
\log\sigma_i &= \beta_{\sigma_{0,sample[i]}} + \beta_{\sigma_1} CategoryPair \\
\beta_{\sigma_1} &\sim Normal(0, 2) \\
\beta_{\sigma_{0,sample}} &\sim Cauchy_+(0, 0.3) \\
\nu &\sim Gamma(2, 0.1)
\end{aligned}
$$

and, b) different domains (e.g., propensity-general and frequency-smoking).

$$
\begin{aligned}
y_i &\sim StudentT(\nu, \theta_i, \sqrt{se_i^2 + \sigma_i^2}) \\
\theta_i &= \beta_{\theta_{0,sample[i]}} + \beta_{\theta_1} DomainPair \\
\beta_{\theta_1} &\sim Normal(0, 1) \\
\beta_{\theta_{0,sample}} &\sim Cauchy_+(0, 0.3) \\
\log\sigma_i &= \beta_{\sigma_{0,sample[i]}} + \beta_{\sigma_1} DomainPair \\
\beta_{\sigma_1} &\sim Normal(0, 2) \\
\beta_{\sigma_{0,sample}} &\sim Cauchy_+(0, 0.3) \\
\nu &\sim Gamma(2, 0.1)
\end{aligned}
$$

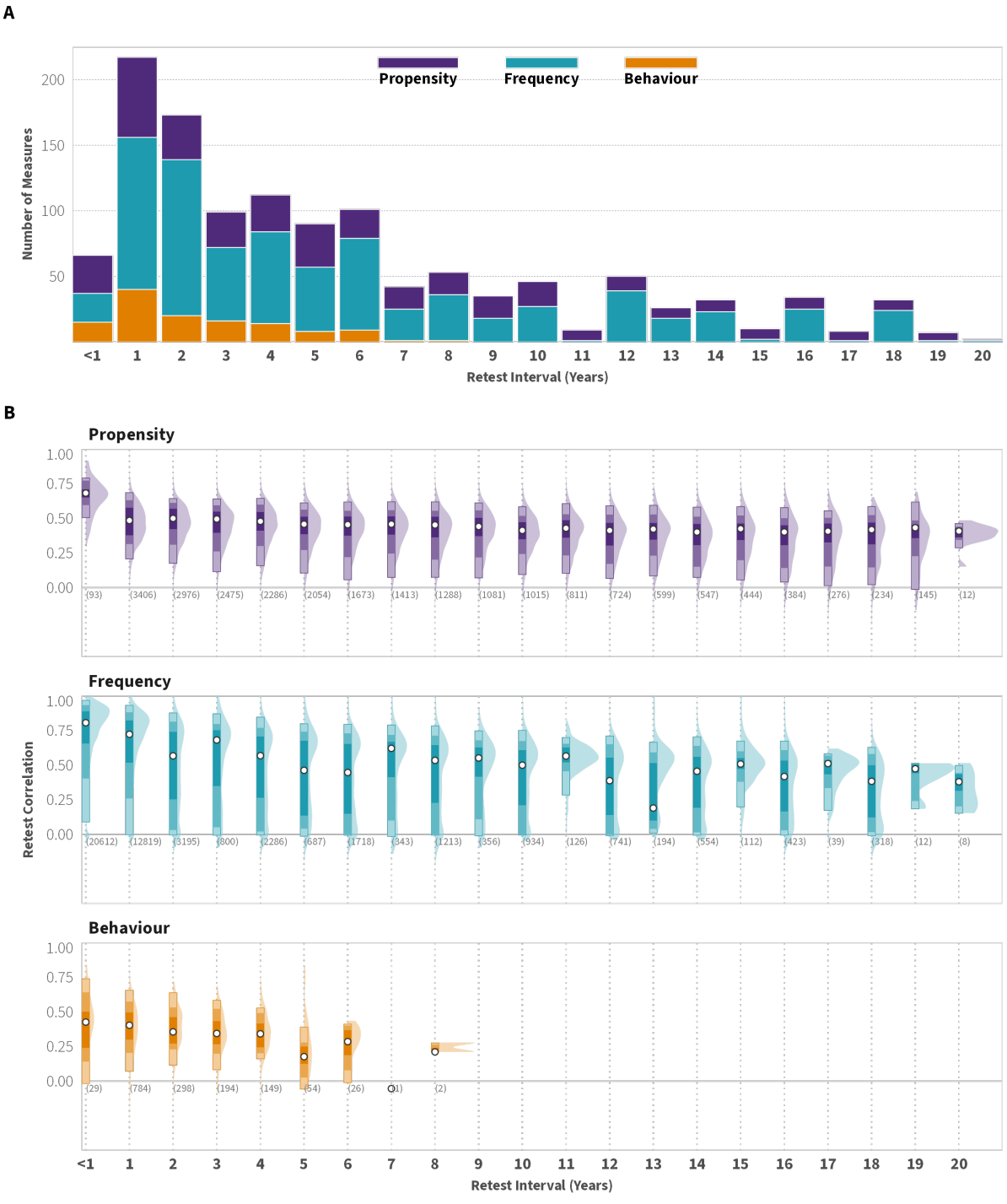
In both meta-regressions we specified predictors for the residual standard deviations, and allowed it to vary across the different levels of the categorical variables. Based on recommendations, in all models, we used weakly informative priors (Williams et al., 2018). Lastly, we back-transformed the results to Spearman's rho for the reporting.

Multiverse Analyses

For brevity and ease of communication, we limited the reporting to a single data set that was the result of a specific set of data pre-processing and processing choices. To communicate transparently about our results and evaluate their robustness (i.e., how sensitive results were to different data processing choices), we repeated our main analyses using different data sets and model specifications (Steegen et al., 2016). On the companion website we describe the different steps and choices that were available when constructing and analysing the data, and include a visual summary of the alternative results (Hall et al., 2022).

Figure S1

Overview of temporal stability measures and correlations. A) The number of measures by category (propensity, frequency, behaviour) and retest interval. B) Distributions of retest correlations as a function of retest interval for the different measure categories (propensity, frequency, behaviour).



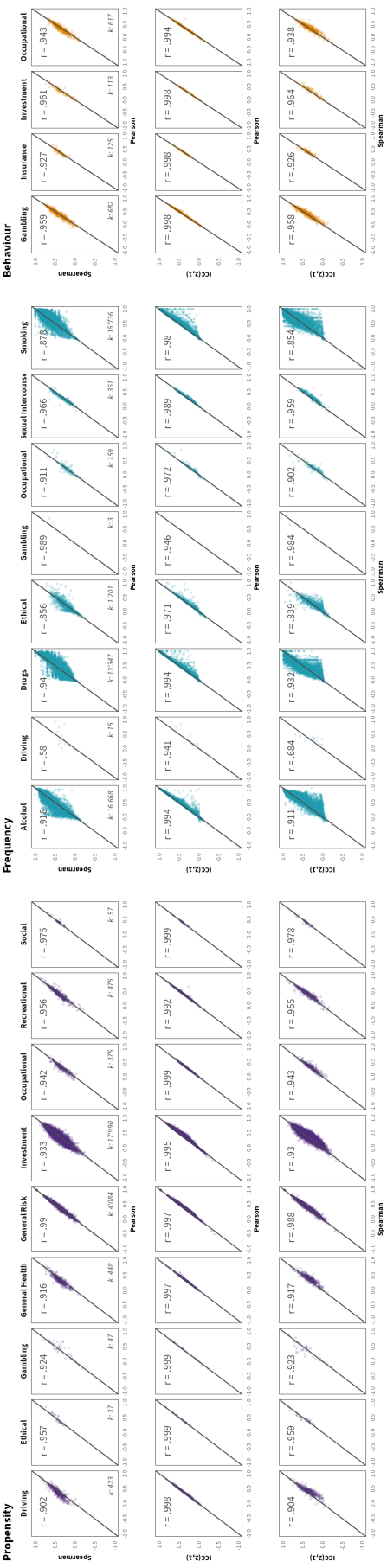


Figure S2

Scatter plots of test-retest correlations calculated using Pearson's r, Spearman's rho, or ICC(2,1).

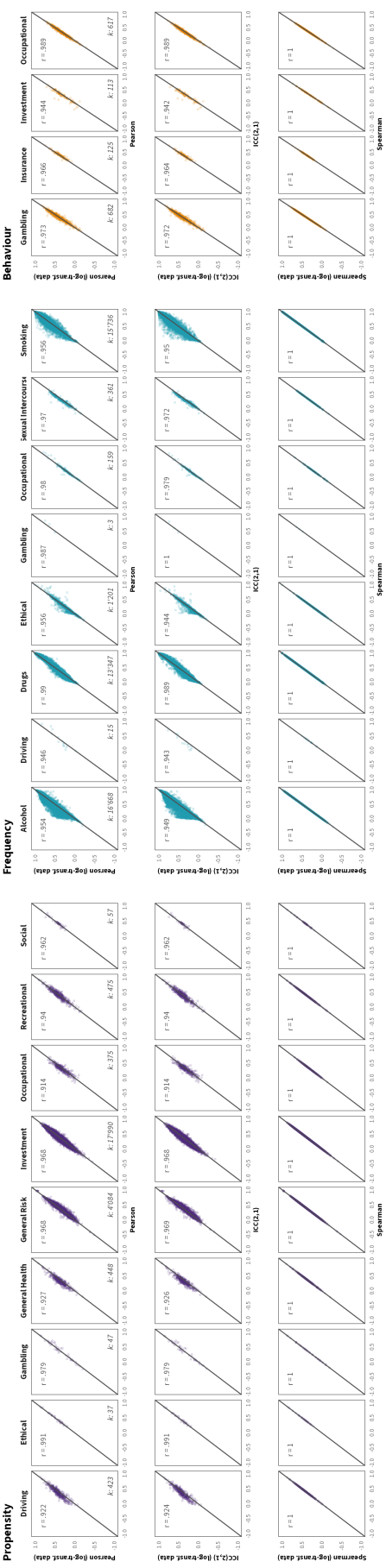


Figure S3

Scatter plots of different test-retest metrics calculated using either log-transformed or non-transformed data

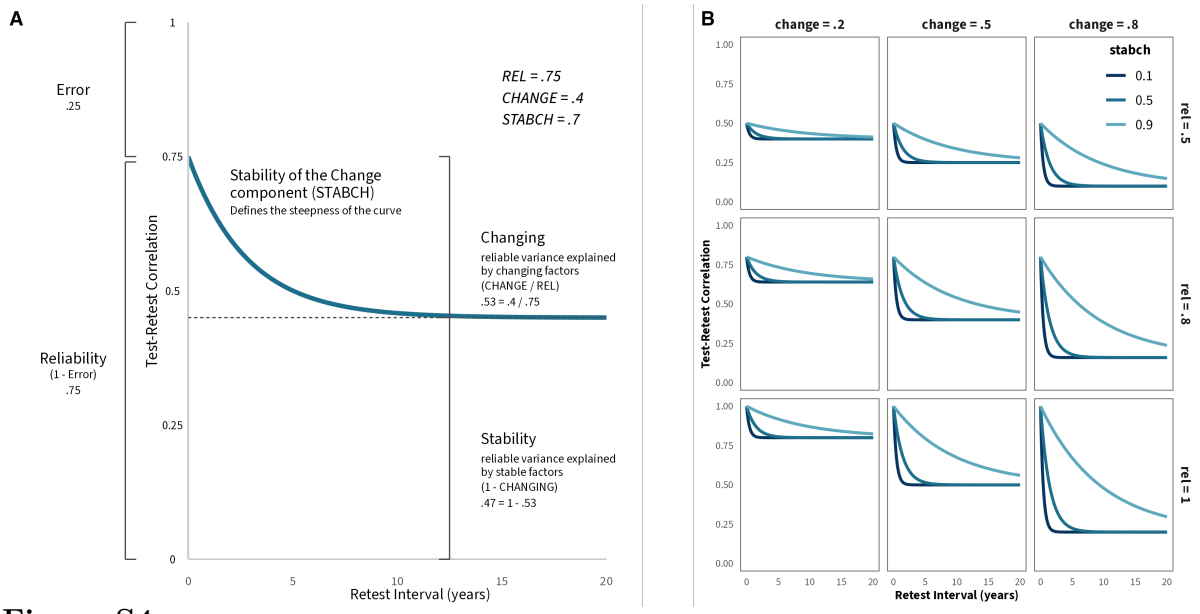


Figure S4

Depiction of the Meta-Analytic Stability and Change model (MASC). A) Visual depiction of temporal stability curve for major personality traits as estimated by Anusic and Schimmack (2016). B) Examples of different parameterisations of MASC.

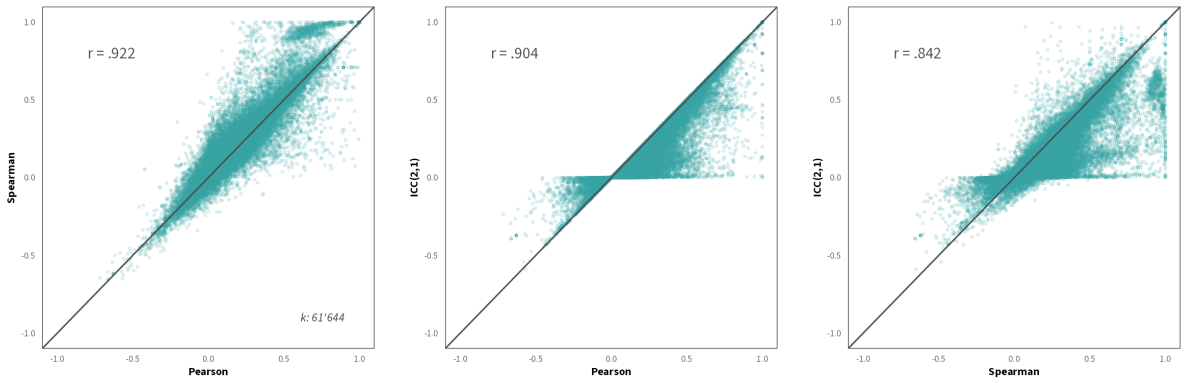


Figure S5

Scatter plots of inter-correlations computed using Pearson's r, Spearman's rho, or ICC(2,1).

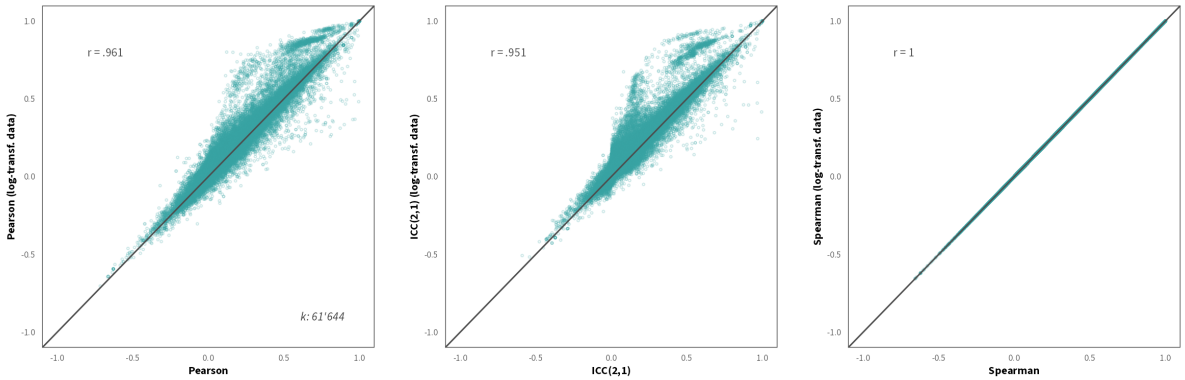


Figure S6

Scatter plots of different inter-correlation metrics calculated using either log-transformed or non-transformed data

Figure S7

Expected values of the posterior predictive distribution (mean, 50%, 80%, and 95% HDI) of Meta-Analytic Stability and Change model (MASC) parameters and test-retest correlations for propensity measures of risk preference in the general (k = 1,732), investment (k = 1,080), and driving (k = 196) domains. Left: Predicted values of the Reliability, Change, and Stability of Change parameters, split by domain, age group and gender. Right: Predicted test-retest correlations as a function of time for different age groups (upper panels) and as a function of age for different retest intervals (lower panels).

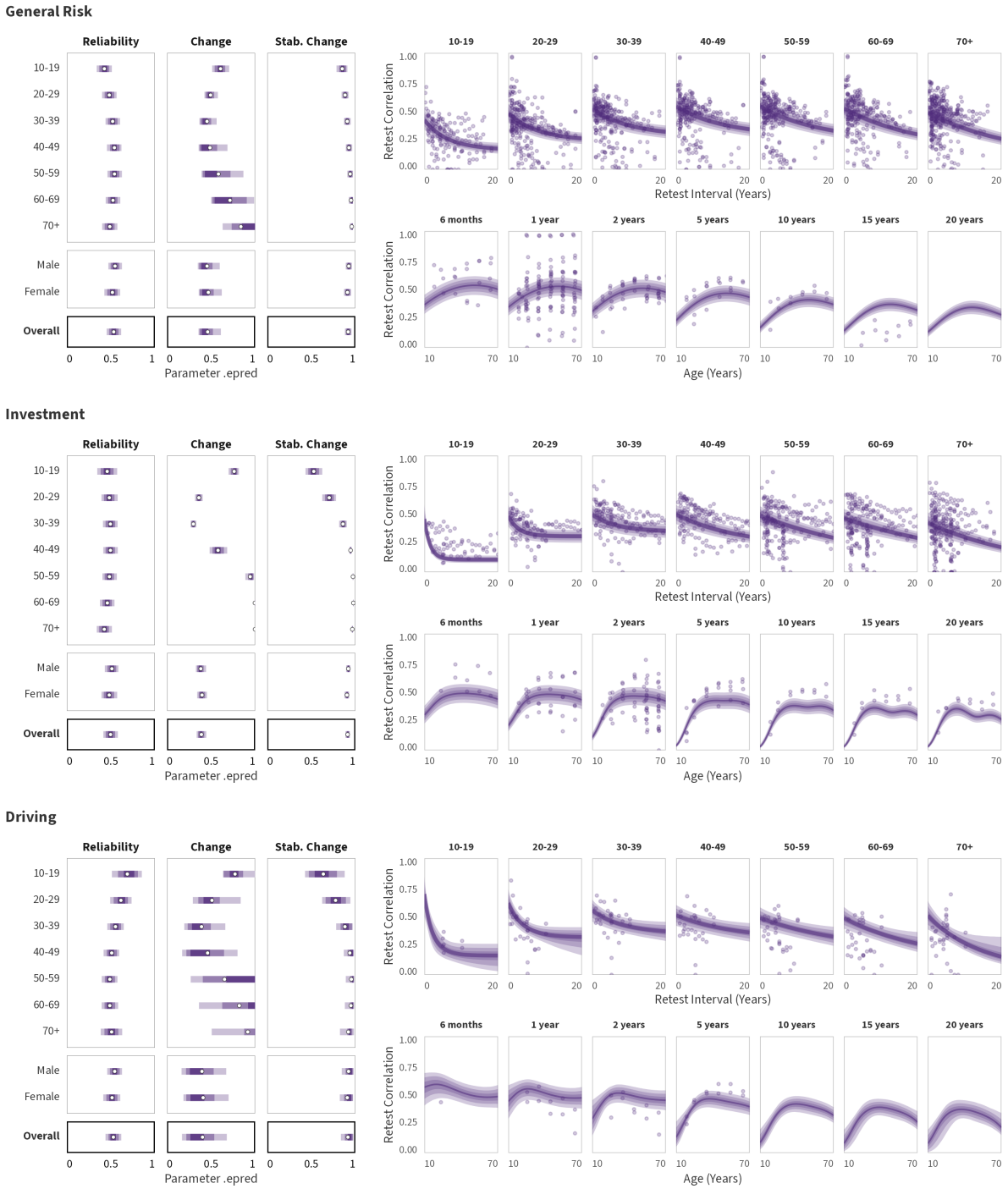


Figure S8

Expected values of the posterior predictive distribution (mean, 50%, 80%, and 95% HDI) of Meta-Analytic Stability and Change model (MASC) parameters and test-retest correlations for propensity measures of risk preference in the ethical (k = 21), gambling (k = 38), and general health (k = 209) domains. Left: Predicted values of the Reliability, Change, and Stability of Change parameters, split by domain, age group and gender. Right: Predicted test-retest correlations as a function of time for different age groups (upper panels) and as a function of age for different retest intervals (lower panels).

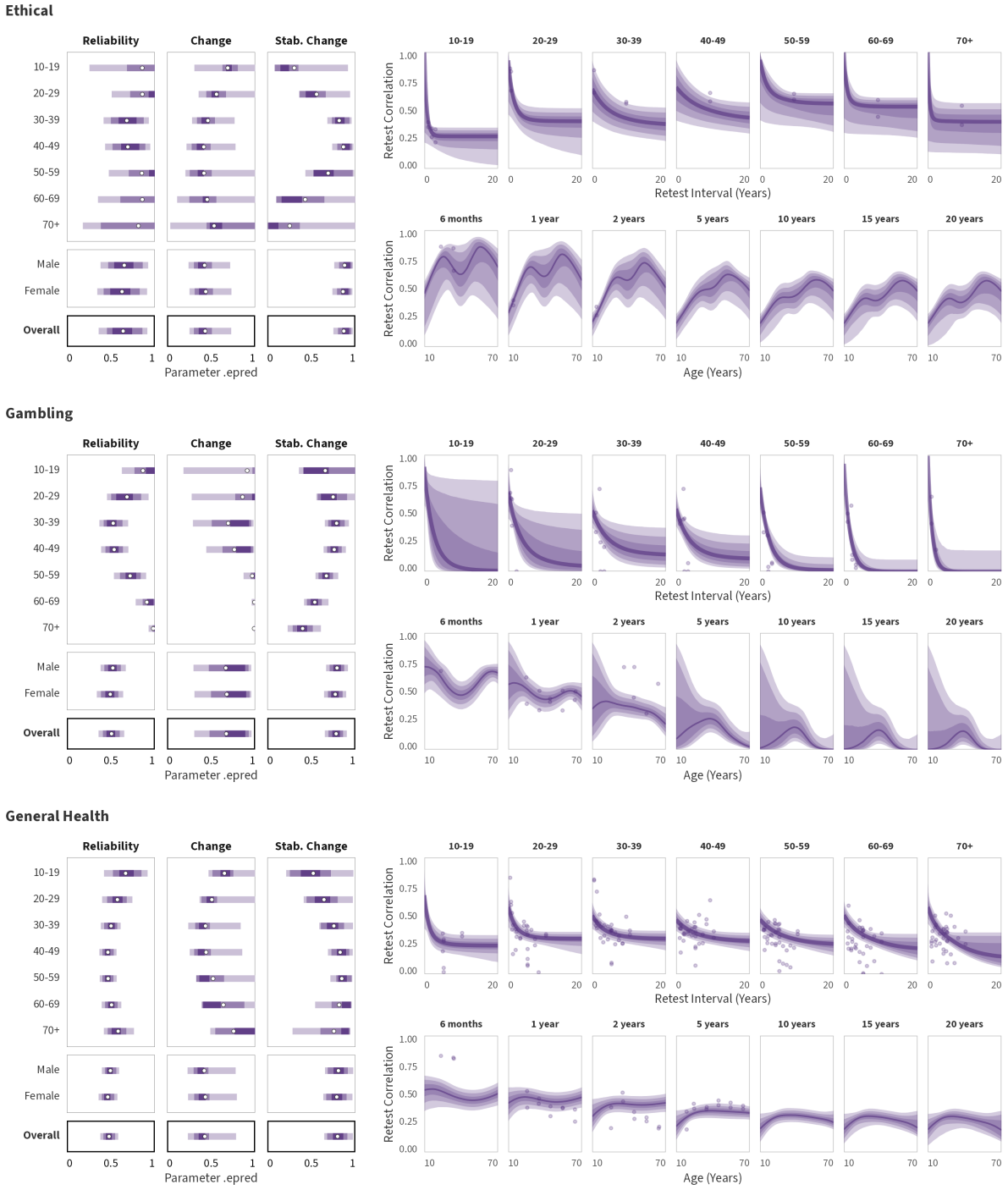


Figure S9

Expected values of the posterior predictive distribution (mean, 50%, 80%, and 95% HDI) of Meta-Analytic Stability and Change model (MASC) parameters and test-retest correlations for propensity measures of risk preference in the occupational (k = 181), recreational (k = 198), and social (k = 51) domains. Left: Predicted values of the Reliability, Change, and Stability of Change parameters, split by domain, age group and gender. Right: Predicted test-retest correlations as a function of time for different age groups (upper panels) and as a function of age for different retest intervals (lower panels).

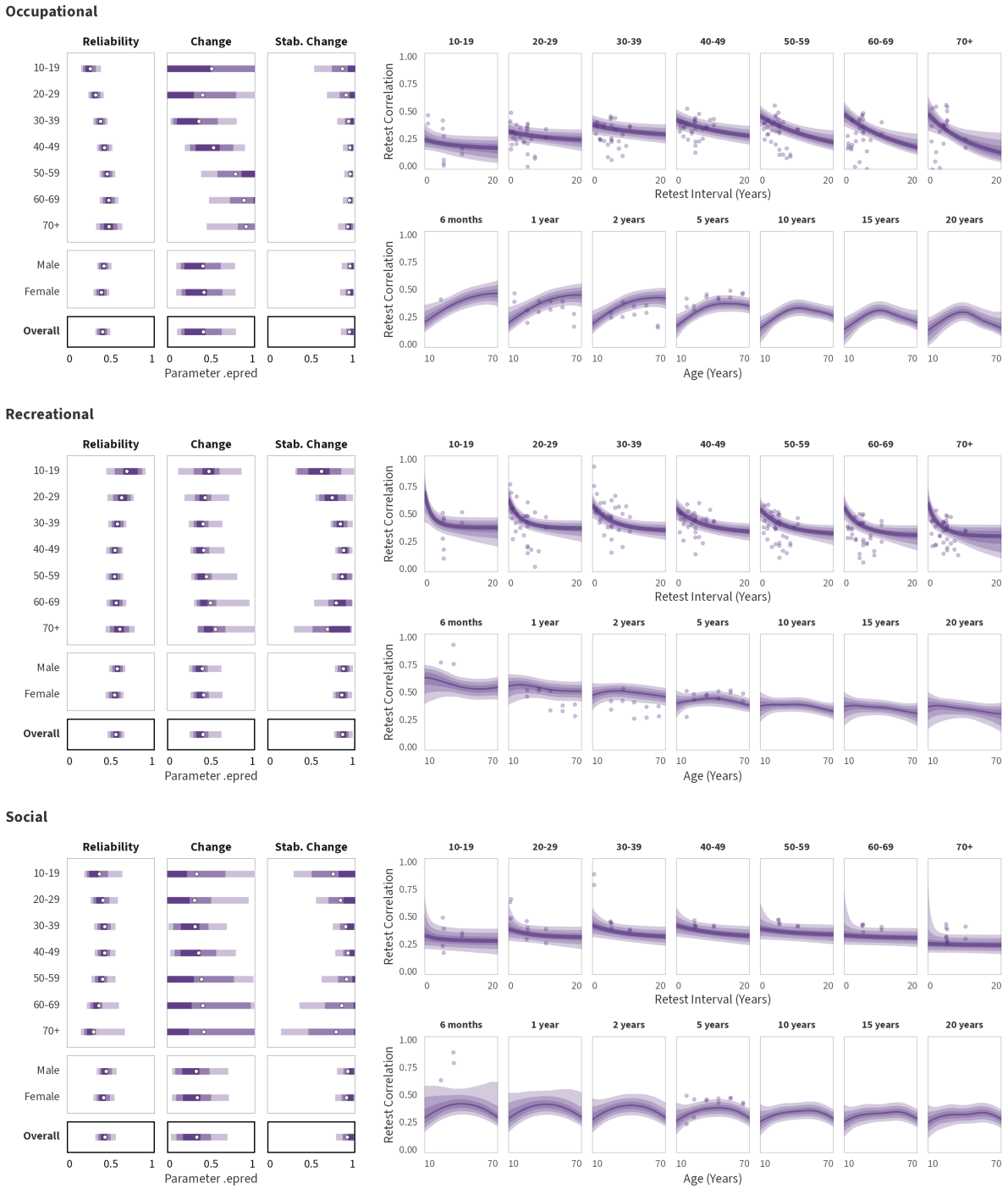


Figure S10

Expected values of the posterior predictive distribution (mean, 50%, 80%, and 95% HDI) of Meta-Analytic Stability and Change model (MASC) parameters and test-retest correlations for frequency measures of risk preference in the alcohol (k = 1,609), driving (k = 15), drugs (k = 223), and ethical (k = 92) domains. Left: Predicted values of the Reliability, Change, and Stability of Change parameters, split by domain, age group and gender. Right: Predicted test-retest correlations as a function of time for different age groups (upper panels) and as a function of age for different retest intervals (lower panels).

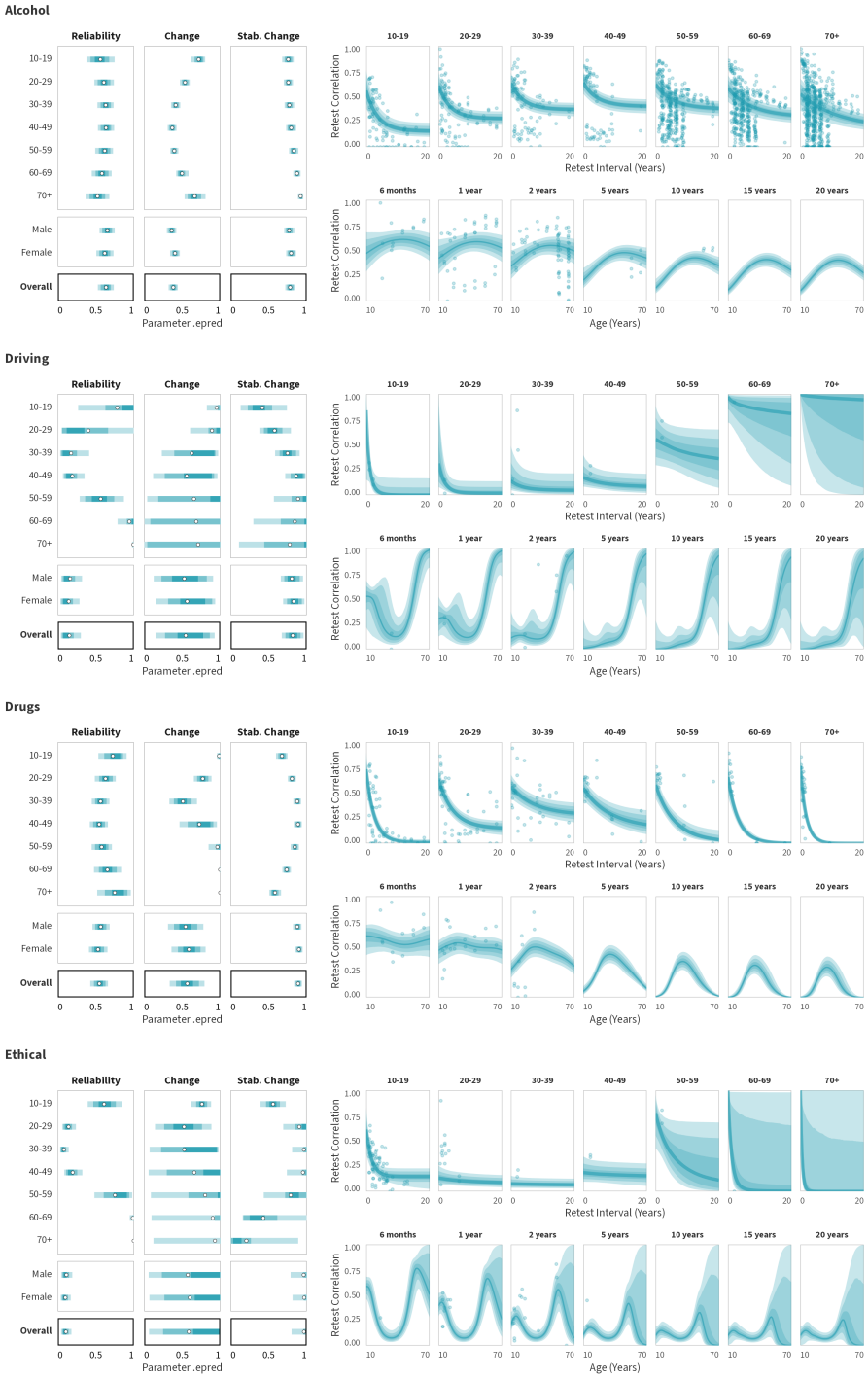


Figure S11

Expected values of the posterior predictive distribution (mean, 50%, 80%, and 95% HDI) of Meta-Analytic Stability and Change model (MASC) parameters and test-retest correlations for frequency measures of risk preference in the smoking (k = 1,637), sexual intercourse (k = 82), gambling (k = 3), and occupational (k = 17) domains. Left: Predicted values of the Reliability, Change, and Stability of Change parameters, split by domain, age group and gender. Right: Predicted test-retest correlations as a function of time for different age groups (upper panels) and as a function of age for different retest intervals (lower panels).

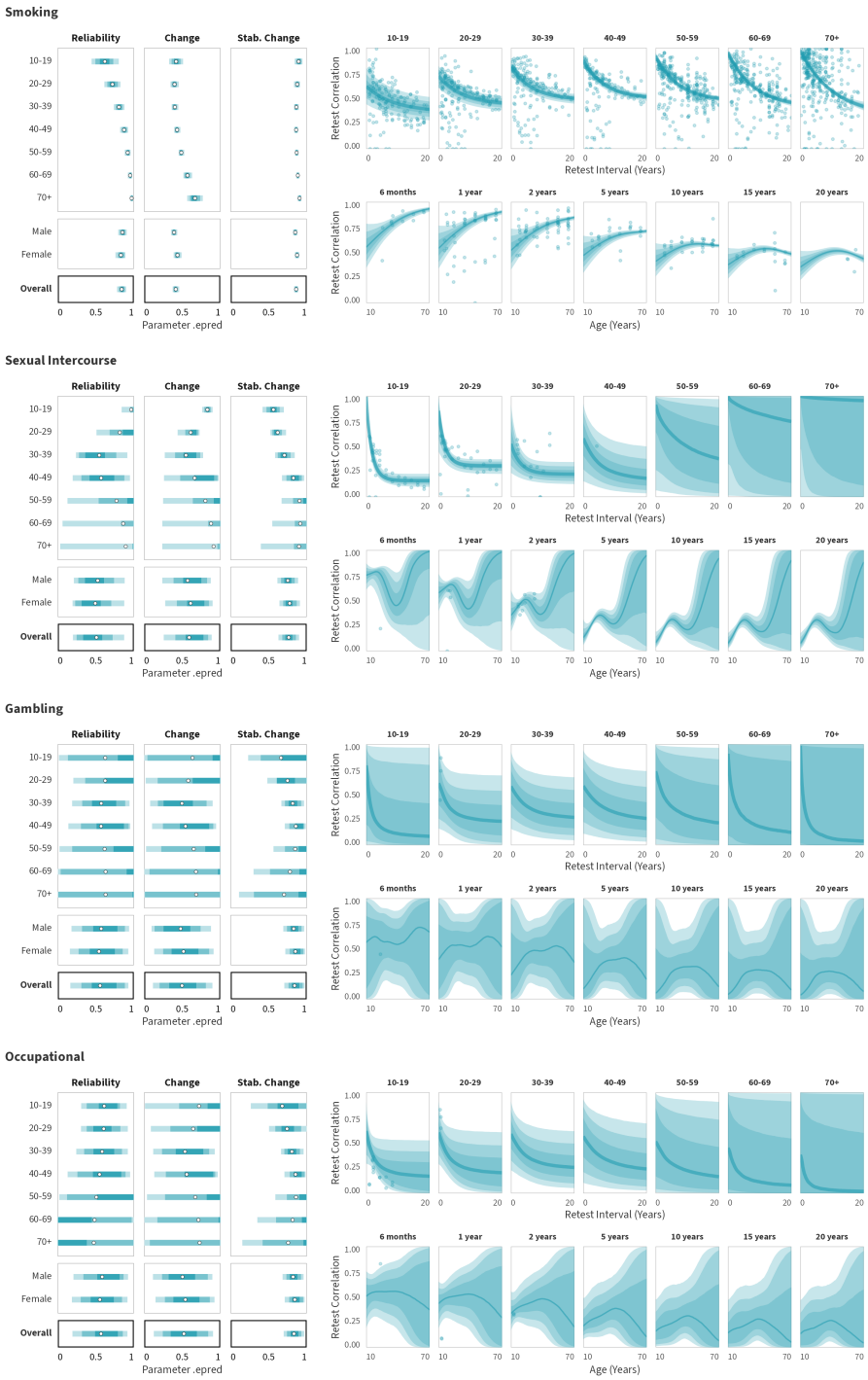


Figure S12

Expected values of the posterior predictive distribution (mean, 50%, 80%, and 95% HDI) of Meta-Analytic Stability and Change model (MASC) parameters and test-retest correlations for propensity measures of risk preference in the investment (k = 108), occupational (k = 227), gambling (k = 197), and insurance (k = 80), domains. Left: Predicted values of the Reliability, Change, and Stability of Change parameters, split by domain, age group and gender. Right: Predicted test-retest correlations as a function of time for different age groups (upper panels) and as a function of age for different retest intervals (lower panels).

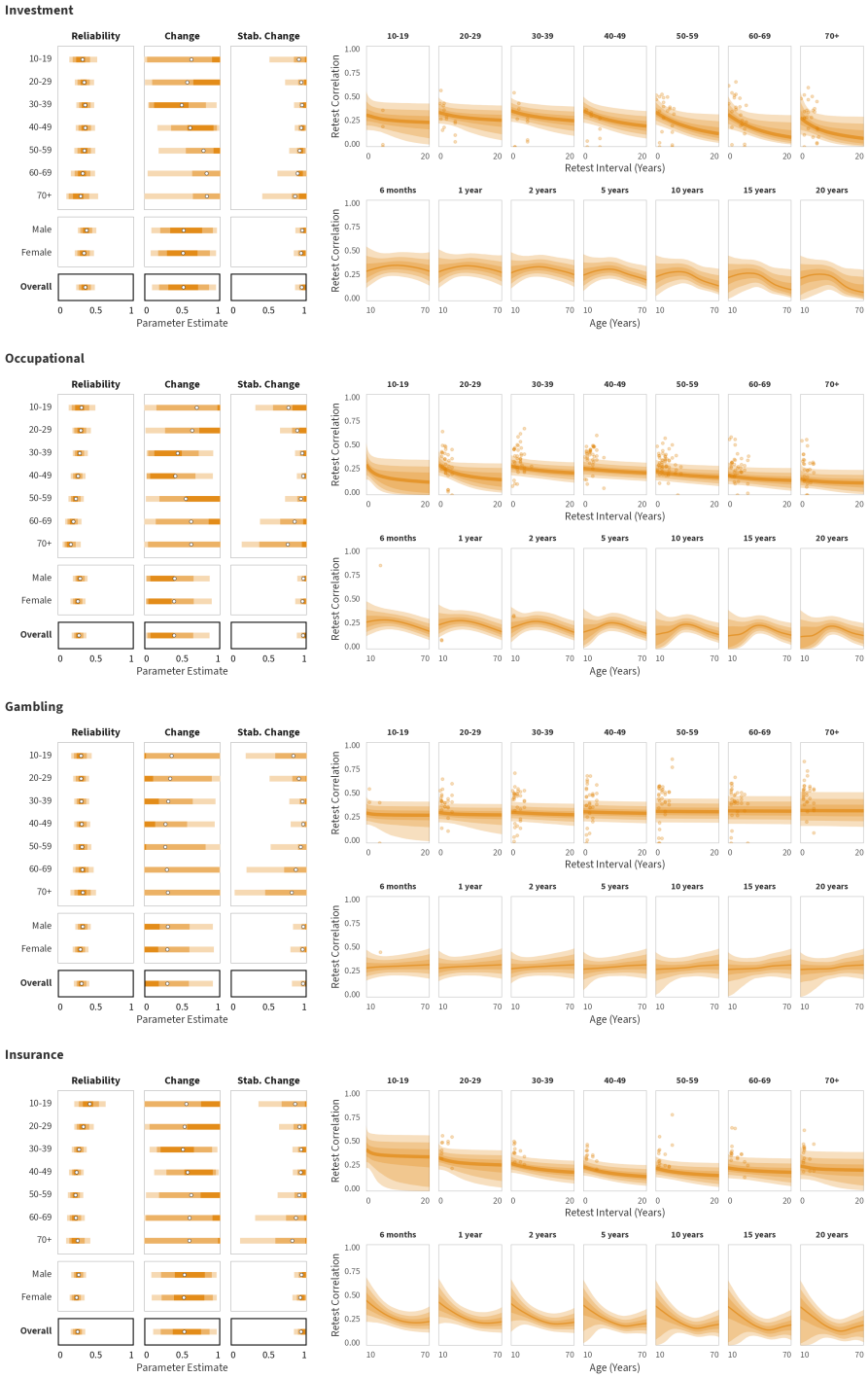


Figure S13

Expected values of the posterior predictive distribution (mean, 50%, 80%, and 95% HDI) of Meta-Analytic Stability and Change model (MASC) parameters and test-retest correlations for personality (k = 226), affect (k = 101), life satisfaction (k = 426), and self-esteem (k = 196). Left: Predicted values of the Reliability, Change, and Stability of Change parameters, split by domain, age group and gender. Right: Predicted test-retest correlations as a function of time for different age groups (upper panels) and as a function of age for different retest intervals (lower panels).

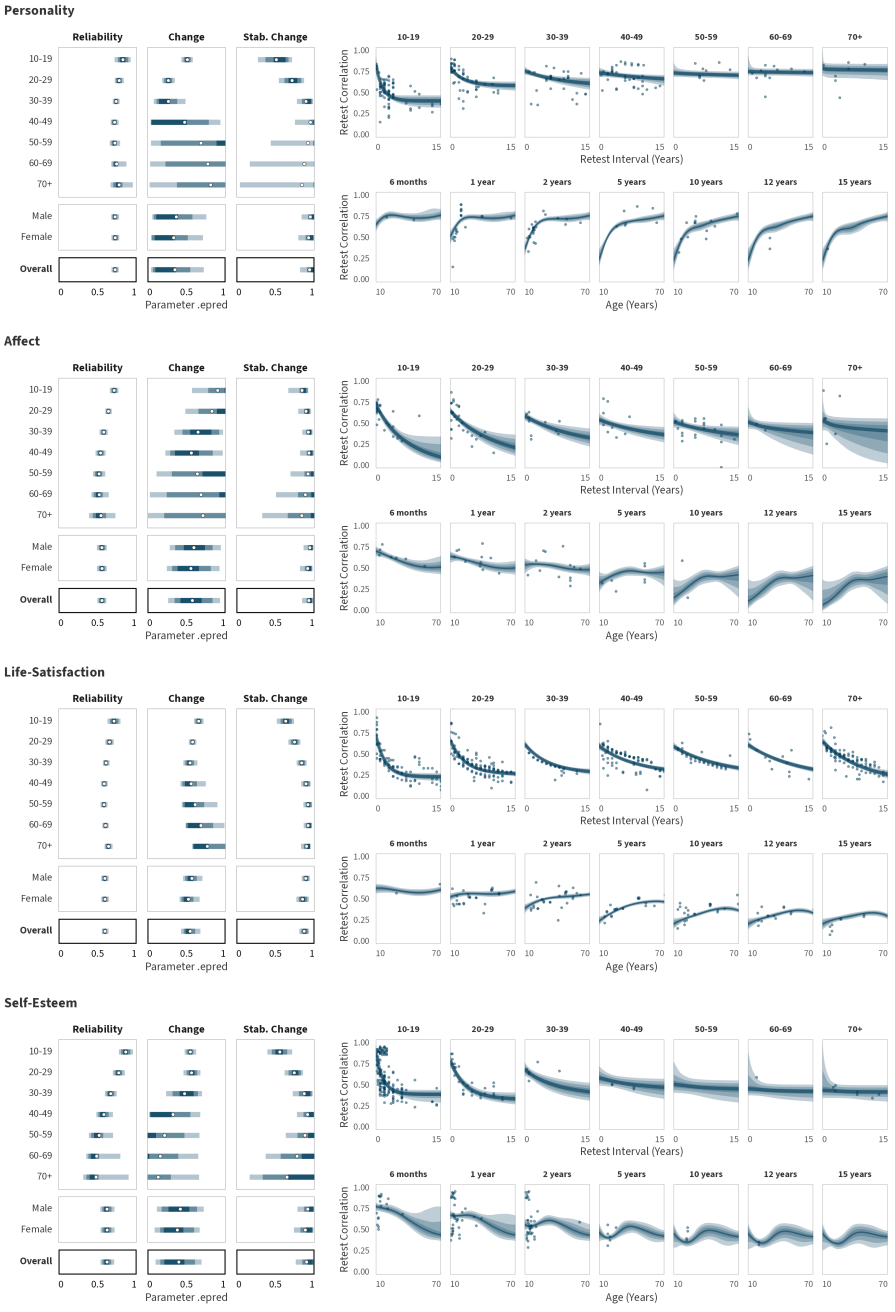


Figure S14

Convergence of risk preference measures. Distributions of inter-correlations between different risk preference measures at the same measurement occasion ($k = 61'644$), split by category-domain pairs (A), and category pairs (B).

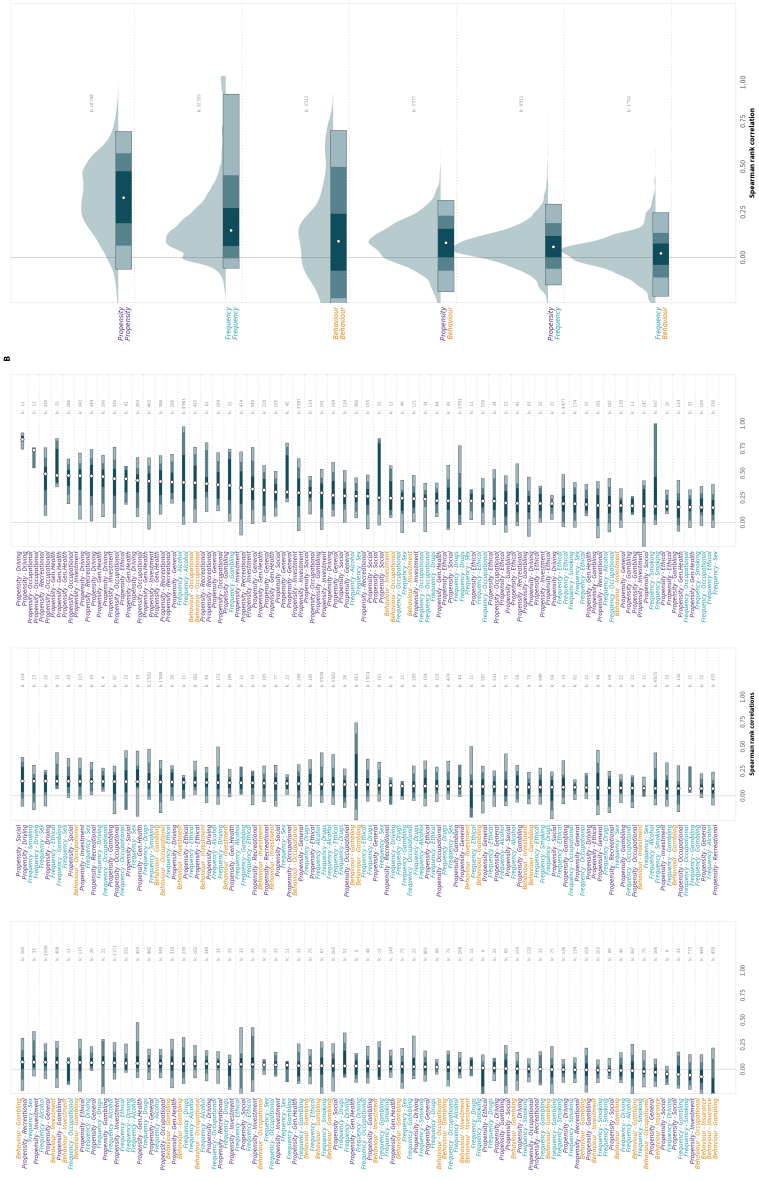


Table S1

Search terms used to identify risk preference measures

Search terms
risk ; attitude ; loss/losing/lose ; excit* ; danger* ; avers* ; chance* ; certain ; safe* ; fear* ;
adventure/venture ; impuls* ; prefer* ; careful
driv* ; car ; fast ; speed ; motor* ; traffic
vandal* ; damage ; cheat ; police ; convict* ; arrest* ; gun ; weapon ; shoot/shot ; troubl*;
stole/steal ; lie ; crim* ; delinquen*; ; aggressive ; fight* ;
assault ; violen* ; injur* ; bully ; affair ; *faith*
finan* ; gambl* ; lottery ; coin ; invest* ; stocks ; bet* ;
casino ; fund ; poker ; trad* ; shares ; bonds
health ; drug ; alcohol ; smok* ; drink* ; cigarette ; drunk ; intox* ; marijuana/cannabis ; heroin ;
meth* ; cocaine ; stimulant ; ecstasy ; hallucinogen ;
tobacco ; wine ; liquor ; spirit ; beer/pint; unprotected sex/intercourse ; contracept*
occupation ; career ; job ; self-employ* ; employ* ; business ; work* ; entrepreneur
extreme ; sport ; bar/pub ; night* ; mountain* ; skydiving ; bunjee ; ski ; climb* ; race
stranger ; trust

Table S2

Overview of panels screened

Panel/Sample	Status	Reason for exclusion, if applicable
Adema, Nikolka, Poutvaara Sunde (2022); Economics Letters (ANPS) - Czeck Republic sample	Incl.	
Adema, Nikolka, Poutvaara Sunde (2022); Economics Letters (ANPS) - India sample	Excl.	Small sample size
Adema, Nikolka, Poutvaara Sunde (2022); Economics Letters (ANPS) - Mexico sample	Excl.	Small sample size
Adema, Nikolka, Poutvaara Sunde (2022); Economics Letters (ANPS) - Spain sample	Incl.	
American Life Panel (ALP)	Incl.	
American National Election Studies (ANES)	Excl.	Does not include propensity item (that meets criteria)
Americans' Changing Lives study (ACL)	Excl.	Does not include propensity item (that meets criteria)
Basel-Berlin Risk Study (BBRS) - Basel (From Frey et al., 2017 Science Advances)	Incl.	
Basel-Berlin Risk Study (BBRS) - Berlin (From Frey et al., 2017 Science Advances)	Incl.	
Berlin Aging Study (BASE)	Excl.	Restricted data access
Berlin Aging Study-II (BASE-II)	Excl.	Restricted data access
British Election Study 2005-2009 (BES05)	Incl.	
British Election Study 2014-2023 (BES14)	Incl.	
Bundesbank - Panel of Household Finances (PHF)	Incl.	
Bundesbank - Survey on Consumer Expectations	Excl.	Propensity item not asked repeatedly
Bundesbank Online Panel – Haushalte (BOP-HH)	Excl.	Propensity item not asked repeatedly
California Families Project (CFP)	Excl.	Restricted data access
Canadian Longitudinal Study on Aging (CLSA)	Excl.	Does not include propensity item (that meets criteria)
Cape Area Panel Study (CAPS)	Excl.	Propensity item not asked repeatedly
China Health and Retirement Longitudinal Survey (CHARLS)	Excl.	Does not include propensity item (that meets criteria)
Cognition and Aging in the USA	Excl.	Propensity item not asked repeatedly
Collaborative Studies on the Genetics of Alcoholism (COGA)	Excl.	Does not include propensity item (that meets criteria)
Costa Rican Longevity and Healthy Aging Study (CRELES)	Excl.	Does not include propensity item (that meets criteria)
Crime in the Modern City. A Longitudinal Study of Juvenile Delinquency in Münster (CMC)	Incl.	
DNB Household Survey (DHS)	Incl.	
Drichoutis Vassilopoulos (2021); Journal of Economics Management Strategy	Incl.	
Einstein Aging Study (EAS)	Excl.	Restricted data access
English Longitudinal Study of Ageing (ELSA)	Excl.	Propensity item not asked repeatedly
Enkavi et al., 2019 PNAS	Incl.	
Financial Crisis: A Longitudinal Study of Public Response (FICR)	Excl.	Does not include propensity item (that meets criteria)
Fragile Families and Child Wellbeing Study (FFCWS)	Excl.	Does not include propensity item (that meets criteria)
General Social Survey Panel (GSSP)	Excl.	Does not include propensity item (that meets criteria)
German Internet Panel (GIP)	Incl.	
German Longitudinal Election Study (GLES) - Panel 2016-2021	Incl.	

Table S2 cont.

Panel/Sample	Status	Reason for exclusion, if applicable
German Longitudinal Election Study (GLES) - Long-term Online Tracking, Cumulation	Excl.	Does not include propensity item (that meets criteria)
German Longitudinal Election Study (GLES) - Long-term Panel 2002-2005-2009	Excl.	Does not include propensity item (that meets criteria)
German Longitudinal Election Study (GLES) - Long-term Panel 2005-2009-2013	Excl.	Does not include propensity item (that meets criteria)
German Longitudinal Election Study (GLES) - Long-term Panel 2009-2013-2017	Excl.	Does not include propensity item (that meets criteria)
German Longitudinal Election Study (GLES) - Short-term Campaign Panel 2009	Excl.	Propensity item not asked repeatedly
German Longitudinal Election Study (GLES) - Short-term Campaign Panel 2013	Excl.	Does not include propensity item (that meets criteria)
German Longitudinal Election Study (GLES) - Short-term Campaign Panel 2013-2017 (repeatedly questioned respondents)	Incl.	
German Longitudinal Election Study (GLES) - Short-term Campaign Panel 2017	Excl.	Propensity item not asked repeatedly
Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI)	Excl.	Does not include propensity item (that meets criteria)
Health Retirement Survey (HRS)	Incl.	
Health Retirement Survey: Cognitive Economics Project (CogEcon)	Incl.	
Health, Aging, and Retirement in Thailand (HART)	Excl.	Does not include propensity item (that meets criteria)
Healthy Ageing in Scotland (HAGIS)	Excl.	Only W1/Pilot data available
High School and Beyond (HSB)	Excl.	Does not include propensity item (that meets criteria)
Household Finance and Consumption Survey (HFCS)	Excl.	Does not include propensity item (that meets criteria)
Household, Income and Labour Dynamics in Australia (HILDA)	Incl.	
Indonesia Family Life Survey (IFLS)	Excl.	Does not include propensity item (that meets criteria)
Interdisciplinary Longitudinal Study of Adult Development (ILSE and ILSE.Y)	Excl.	Restricted data access
Japan Household Panel Survey (JHPS)	Excl.	Does not include propensity item (that meets criteria)
Japanese Study of Aging and Retirement (JSTAR)	Excl.	Does not include propensity item (that meets criteria)
Korean Labour Income Panel Survey	Excl.	Does not include propensity item (that meets criteria)
Korean Longitudinal Study of Aging (KLoSA)	Excl.	Does not include propensity item (that meets criteria)
Life in Kyrgyzstan Study (LIKS)	Incl.	
Longitudinal Aging Study in India (LASI)	Excl.	Only W1/Pilot data available
Longitudinal Aging Study of Amsterdam (LASA)	Excl.	Does not include propensity item (that meets criteria)
Longitudinal Internet studies for the Social Sciences (LISS)	Excl.	Propensity item not asked repeatedly
Longitudinal Study of American Youth (LSAY)	Excl.	Does not include propensity item (that meets criteria)
Longitudinal Study of Australian Children (LSAC)	Excl.	Propensity item not asked repeatedly
Longitudinal Study of Violence Against Women - Men Sample (LSVAW-M)	Incl.	
Longitudinal Study of Violence Against Women - Women Sample (LSVAW-W)	Incl.	
Longitudinal Surveys of Australian Youth (LSAY)	Excl.	Does not include propensity item (that meets criteria)
Lothian Birth Cohort 1936	Excl.	Restricted data access
Malaysia Ageing and Retirement Survey (MARS)	Excl.	Only W1/Pilot data available
Medical Expenditure Panel Survey (MEPS)	Incl.	
Mexican Family Life Survey (MxFLS)	Excl.	Propensity item not asked repeatedly
Mexican Health and Aging Study (MHAS)	Excl.	Does not include propensity item (that meets criteria)

Table S2 cont.

Panel/Sample	Status	Reason for exclusion, if applicable
Midlife in Japan (MIDJA)	Incl.	
Midlife in the United States (MIDUS) - Milwaukee Dample	Excl.	Restricted data access
Midlife in the United States (MIDUS) - Project 1 Sample	Incl.	
Millennium Cohort Study (MCS)	Excl.	Propensity item not asked repeatedly
Monitoring the Future: Restricted-Use Panel Data	Excl.	Restricted data access
National Health and Nutrition Examination Survey (NHANES)	Excl.	Does not include propensity item (that meets criteria)
National Income Dynamics Study (NIDS)	Excl.	Does not include propensity item (that meets criteria)
National Longitudinal Study of Adolescent to Adult Health (Add Health)	Incl.	
National Longitudinal Survey of Youth 1979 (NLSY79)	Excl.	Propensity item not asked repeatedly
National Longitudinal Survey of Youth 1979 Child and Young Adult (NLSY79-CYA)	Incl.	
National Longitudinal Survey of Youth 1997 (NLSY97)	Excl.	Propensity item not asked repeatedly
National Social Life, Health, and Aging Project (NSHAP)	Incl.	
National Survey of Families and Households (NSFH)	Excl.	Does not include propensity item (that meets criteria)
New Zealand Health, Work and Retirement Study	Excl.	Does not include propensity item (that meets criteria)
Nießen et al. (2020) . GESIS Instrument	Excl.	Small sample size
Northern Ireland Cohort for the Longitudinal Study of Ageing (NICOLA)	Excl.	Propensity item not asked repeatedly
Origin of Variance in the Oldest-Old: Octogenerian Twins (Octo-Twin)	Excl.	Does not include propensity item (that meets criteria)
Panel Study of Income Dynamics (PSID)	Excl.	Propensity item not asked repeatedly
Panel Survey of Consumer Finances 1983-1989	Excl.	Cannot match respodents
Panel Survey of Consumer Finances 2007-2009	Excl.	Does not include propensity item (that meets criteria)
Parenting Across Cultures	Excl.	Does not include propensity item (that meets criteria)
Preference Parameters Study - India (rural area) (GCOE - IN Rural)	Excl.	Does not include propensity item (that meets criteria)
Preference Parameters Study - India (urban area) (GCOE - IN)	Incl.	
Preference Parameters Study - China (urban area) (GCOE - CN)	Incl.	
Preference Parameters Study - Japan (GCOE - JP)	Incl.	
Preference Parameters Study - United States of America (GCOE - USA)	Incl.	
Public Opinion and the Syrian Crisis in Three Democracies	Excl.	Propensity item not asked repeatedly
Risky decision and happiness task: The Great Brain Experiment smartphone app	Excl.	Does not include propensity item (that meets criteria)
Rochester Adult Longitudinal Study (RALS)	Excl.	Does not include propensity item (that meets criteria)
Rural-Urban Migration in China and Indonesia: CHINA	Excl.	Missing documentation
Rural-Urban Migration in China and Indonesia: INDONESIA	Excl.	Propensity item not asked repeatedly
Russian Longitudinal Monitoring Survey (RLMS-HSE)	Excl.	Restricted data access
Screening Across the Lifespan Twin Study: the Younger (SALTY)	Excl.	Restricted data access
Seattle Longitudinal Study (SLS)	Excl.	Does not include propensity item (that meets criteria)
Socio-Economic Panel Study - Core (SOEP-Core SOEP-CoV)	Incl.	
Socio-Economic Panel Study Retest (SOEP-Retest)	Excl.	Small sample size
Sparen und Altersvorsorge in Deutschland (SAVE)	Incl.	

Table S2 cont.

Panel/Sample	Status	Reason for exclusion, if applicable
Steiner et al., (2020); Decision	Excl.	Small sample size
Studies Incl. in Enkavi et al. (2019 PNAS) meta-analyis	Excl.	No open access data
Studies Incl. in Mata et al. (2018 JEP) meta-analysis	Excl.	No open access data
Study to Assess Risk and Resilience in Servicemembers — Longitudinal Study (STARRS)	Excl.	Propensity item not asked repeatedly
Survey of Consumer Expectations (SCE)	Excl.	Propensity item not asked repeatedly
Survey of Health, Ageing and Retirement in Europe (SHARE) (Excluding the following countries: Bulgaria, Croatia, Cyprus, Finland, Greece, Hungary, Latvia, Lithuania, Luxemburg, Malta, Poland, Portugal, Romania, Slovak Republic)	Excl.	Propensity item not asked repeatedly
Survey of Health, Ageing and Retirement in Europe (SHARE) (Including the following countries: Austria, Belgium, Czech_Rep, Denmark, Estonia, France, Germany, Israel, Italy, Netherlands, Slovenia, Spain, Sweden, Switzerland)	Incl.	
Swedish Adoption/Twin Study of Aging (SATSA)	Excl.	Small sample size
Swiss Household Panel (SHP)	Excl.	Propensity item not asked repeatedly
Thailand Vietnam Socio Economic Panel (TVSEP)	Excl.	Does not include propensity item (that meets criteria)
The Brazilian Longitudinal Study of Aging (ELSI-Brazil)	Excl.	Does not include propensity item (that meets criteria)
The Irish LongituDinal Study on Ageing (TILDA)	Excl.	Does not include propensity item (that meets criteria)
Tracking Adolescents' Individual Lives Survey (TRAILS)	Excl.	Does not include propensity item (that meets criteria)
TwinLife	Incl.	
Twins of Early Development Study (TEDS)	Excl.	Propensity item not asked repeatedly
UK Biobank	Excl.	Does not include propensity item (that meets criteria)
UK Household Longitudinal Survey + British Household Panel Survey (USOC)	Excl.	Propensity item not asked repeatedly
UK Household Longitudinal Survey Innovation Panel (USOC_IP)	Incl.	
Ukrainian Longitudinal Monitoring Survey (ULMS)	Incl.	
Understanding America Study (UAS)	Incl.	
VA Normative Aging Study (VA NAS)	Excl.	Does not include propensity item (that meets criteria)
WHO Study on global AGEing and adult health (SAGE)	Excl.	Does not include propensity item (that meets criteria)
Wisconsin Longitudinal Study (WLSG/WLSS)	Excl.	Does not include propensity item (that meets criteria)
Work and Family Life Study	Excl.	Propensity item not asked repeatedly

End of Table

Table S3

Overview of exclusion and inclusion criteria of measures for the analyses, split by measure category

Category	Inclusion	Exclusion	Rationale
All	1. Measures that have been asked to the same respondents across at least two time points.	1. Measures that have been asked only in one wave or only once to the respondents	1. We need responses from a least two time points two compute a test-retest correlation coefficient.
All	2. Measures where the wording and response format remained consistent across at least two time points.	2. Measures that are not consistent across at least two time points	2. Measures need to be the same across waves to accurately measure test-retest correlations
All	3. Measures that include at least 4 response options/values, or is composed of multiple (binary) measures that can be aggregated to calculate an index.	3. Measures that include less than four response options/values (e.g., yes/no, never/sometimes/always).	3. With more response options it is possible to capture more meaningful changes over time .
All	4. Measures that use an ordinal scale, discrete scale (with a clear response range) or are open-ended	4. Measures that use a nominal scale or scales with options that cannot be objectively ranked	4. Can result in subjective interpretations of what a category is and thus reduces response comparability between participants. Further if response options cannot be ranked, this can reduce the accuracy of how the test-retest correlations are computed.
Propensity	1. Measures that ask respondents about recent behaviour.	1. Measures that ask respondents about behaviour that is too far back in time or no longer relevant (e.g., asking adult respondents about their risk propensity as a child).	1. Relies on the recollection of certain events, which can result in inaccuracies. We are not capturing temporal stability based on the responses of actions that are no longer relevant .
Propensity	2. Measures that refer directly to the respondent.	2. Measures that refer to an individual other than the respondent (e.g., partner/spouse, household)	2. Another person's or group's risk preference is not necessarily reflective of the respondent's. Thus, individual changes would not be reflected in the response.
Propensity	3. Measures that can be answered by both women and men	3. Gender-specific measures (e.g., specific behaviour during pregnancy)	3. We want to collect approximately the same amount of responses from both males and females respondents to best explore gender differences.
Propensity	4. Measures that explicitly ask about risk-taking.	4. Measures that ask about ambiguity.	4. Ambiguity preference is shown to differ from risk preference (Levy et al., 2010)
Propensity	5. Measures that can be classified into a general or single life domain (e.g., general, driving, recreational)	5. Measures for which the behaviour cannot be classified into more than one pre-specified domain	5. More accurate comparison across domains
Frequency	1. Measures that ask respondents about recent or ongoing behaviour.	1. Measures that ask respondents about behaviour that is too far in time or no longer relevant (e.g., number of cigarettes smoked before quitting).	1. Relies on the recollection of certain events, which can result in inaccuracies. Asking about behaviours that are no longer taking place in the present can result in inflated correlation coefficients.
Frequency	2. Measures with a clearly specified time frame (e.g., in the last month/week how often...).	2. Measures with no clearly specified time frame or that refer to the course of the respondent's life time or that are dependent on a specific event (e.g., since you were 14 years old).	2. Such questions do not allow a proper comparison between participants as these can result in the subjective interpretation of a time frame or they are dependent on other factors (e.g., current age).

Table S3 cont.

Category	Inclusion	Exclusion	Rationale
Frequency	3. Measures that refer directly to the respondent.	3. Measures that refer to an individual other than the respondent (e.g., partner/spouse, household)	3. Another person's or group's risk preference is not necessarily reflective of the respondent's. Thus, individual changes would not be reflected in the response.
Frequency	4. Measures that use an ordinal scale, discrete scale (with a clear response range) or are open-ended	4. Measures that use a nominal scale or scales than cannot be objectively ranked	4. Can result in subjective interpretations of what a category is and thus reduces response comparability between participants. Further if response options cannot be ranked, this can reduce the accuracy of how the test-retest correlations are computed.
Frequency	5. Measures that include 0 or Never response options	6. Measures that do not include 0 or Never response options	5. It is possible to enter a response for those respondent whom this question does not apply (e.g., non-smokers smoking 0 cigarettes). Additionally, such measures help better capture changes across time (e.g., a frequent smoker at T1 but quits smoking at T2)
Frequency	5. Measures that can be answered by both women and men	6. Gender-specific measures (e.g., specific behaviour during pregnancy)	6. We want to collect the same amount of responses from both males and females respondents to best explore gender differences.
Frequency	6. Measures that can be classified into a single life domain (e.g., smoking, alcohol, driving)	6. Measures for which the behaviour can be classified into more than one life domain	6. More accurate comparison across domains
Behaviour	1. Measures with choices that vary on in terms of probabilities, or that have a clear risk component.	1. Measures with choices that not solely vary in terms of probabilities (e.g. choices dependent on the response of another individual, choices involving a dimension of time).	1. Including measures that vary on other dimensions of the choice options would result in risk preference being confounded by other preferences (e..g, social preference, time preference)
Behaviour	2. Measures with choices that involve a form of monetary outcome or reward.	2. Measures with choices in non-financial contexts with other forms of outcomes	2. Such measures allow a direct comparison to tasks commonly using the economics literature

End of Table

Table S4

Overview of panels included in the analyses

Sample	Country	Collect	Oper.	Domains	N.meas.	N.waves	N.corr	N
ADDHEALTH	U.S.A.	Int.	F, P	Alc., Dri., Dru., Eth., Gen., Sex., Smo.	49	5	379	6,138
ALP	U.S.A.	Onl.	P, B	Gen., Inv., Gam., Occ.	11	18	215	3,180
ANPS-Czech-Republic	Czech Republic	Onl.	P, B	Gen., Inv.	2	2	4	230
ANPS-Spain	Spain	Onl.	P, B	Gen., Inv.	2	2	5	177
BBRS-CH	Switzerland	Lab.	F, P, B	Alc., Dri., Dru., Eth., Gam., Gen., Hea-gen., Inv., Occ., Rec., Sex., Soc.	35	2	35	34
BBRS-DE	Germany	Lab.	F, B, P	Alc., Eth., Sex., Occ., Gam., Dru., Dri., Gen., Inv., Hea-gen., Rec., Soc.	35	2	70	99
BES05	U.K.	Onl.	P	Gen.	1	2	12	3,291
BES14	U.K.	Onl.	P, B	Gen., Gam.	2	4	64	32,982
CMC	Germany	Int.	F, P	Eth., Dru., Occ.	25	4	223	2,017
COGECON	U.S.A.	Int.	P, B	Inv., Gen.	3	4	54	871
DHS	Netherlands	Int.	B, P	Gam., Gen., Inv.	7	30	14,161	10,581
DRICHOUTIS	Greece	Self-adm.	P, B	Gen., Inv.	2	3	10	113
ENKAVI	U.S.A.	Onl.	F, P, B	Alc., Dri., Dru., Eth., Gam., Hea-gen., Rec., Smo., Soc.	19	2	32	68
GCOE-CN	China	Int.	P	Gen.	1	2	10	958
GCOE-IN	India	Int.	P, B	Gen., Gam., Occ.	5	5	49	1,280
GCOE-JP	Japan	Self-adm.	P, B	Gen., Occ., Gam., Ins.	15	12	949	8,040
GCOE-USA	U.S.A.	Self-adm.	P, B	Gen., Occ., Gam., Ins.	15	9	684	7,523
GIP	Germany	Onl.	P	Gen.	1	3	32	2,129
GLES-LT	Germany	Int.	P	Gen.	1	6	130	17,320
GLES-ST	Germany	Onl.	P	Gen.	1	2	12	2,045
HILDA	Australia	Int.	P, F	Inv., Gen., Smo.	4	21	5,976	25,154
HRS-Core	U.S.A.	Int.	F, P, B	Alc., Dri., Gen., Hea-gen., Inv., Occ., Rec., Smo.	15	15	2,376	34,027
LIKS	Kyrgyzstan	Int.	F, P	Alc., Gen., Smo.	8	6	758	10,082
LSVAW-M	U.S.A.	Int.	F, P	Alc., Dru., Eth., Gen., Sex.	26	5	306	650
LSVAW-W	U.S.A.	Int.	F, P	Alc., Dru., Eth., Gen., Sex.	23	5	166	1,394
MEPS	U.S.A.	Int.	P	Gen.	1	34	272	157,599
MIDJA	Japan	Int.	P, F	Gen., Alc.	6	2	58	655
MIDUS-Project1	U.S.A.	Int.	F, P	Alc., Dru., Gen., Eth.	9	3	181	4,357
NLSY79-CYA	U.S.A.	Int.	F, P, B	Alc., Dru., Eth., Gen., Occ., Sex., Smo.	31	17	4,222	8,613
NSHAP	U.S.A.	Int.	F, P	Alc., Gen., Smo.	5	3	86	2,943
PHF	Germany	Int.	P	Inv., Gen.	2	3	56	3,566
SAVE	Germany	Self-adm.	F, P	Alc., Dri., Gam., Hea-gen., Inv., Occ., Rec.	9	10	1,895	3,758
SHARE-Austria	Austria	Int.	F, P	Alc., Inv., Smo.	7	7	148	4,863
SHARE-Belgium	Belgium	Int.	F, P	Alc., Inv., Smo.	7	7	191	6,544
SHARE-Czech-Rep	Czech-Rep	Int.	F, P	Alc., Inv., Smo.	6	6	159	5,673
SHARE-Denmark	Denmark	Int.	F, P	Alc., Inv., Smo.	8	7	183	4,249
SHARE-Estonia	Estonia	Int.	F, P	Alc., Inv., Smo.	6	4	80	6,214
SHARE-France	France	Int.	F, P	Alc., Inv., Smo.	7	7	183	5,593
SHARE-Germany	Germany	Int.	F, P	Alc., Inv., Smo.	7	7	160	5,463
SHARE-Israel	Israel	Int.	F, P	Alc., Inv., Smo.	7	5	68	2,665
SHARE-Italy	Italy	Int.	F, P	Alc., Inv., Smo.	7	7	185	5,251
SHARE-Netherlands	Netherlands	Int.	F, P	Alc., Inv., Smo.	7	5	97	3,796
SHARE-Slovenia	Slovenia	Int.	F, P	Alc., Inv., Smo.	6	4	82	3,729
SHARE-Spain	Spain	Int.	F, P	Alc., Inv., Smo.	7	7	174	6,310
SHARE-Sweden	Sweden	Int.	F, P	Alc., Inv., Smo.	7	7	167	4,869
SHARE-Switzerland	Switzerland	Int.	F, P	Alc., Inv., Smo.	7	7	170	3,442

Table S4 cont.

Sample	Country	Collect	Oper.	Domains	N.meas.	N.waves	N. corr	N
SOEP-Core	Germany	Int.	P, B, F	Dri., Gen., Hea-gen., Inv., Occ., Rec., Smo., Soc.	11	19	3,822	61,611
TWINLIFE	Germany	Int.	F, P	Alc., Dri., Eth., Gen., Occ.	18	3	132	9,035
UAS	U.S.A.	Onl.	F, P, B	Alc., Dru., Gen., Inv., Smo.	13	42	32,710	9,371
ULMS	Ukraine	Int.	F, P, B	Alc., Dri., Gen., Hea-gen., Inv., Occ., Rec., Smo.	21	4	277	8,154
USOC-IP	U.K.	Int.	F, B, P	Alc., Dru., Eth., Gam., Gen., Hea-gen., Inv., Smo.	12	13	493	3,707

End of Table

Notes. Mode of data collection: Onl(ine), Self-Adm(inistered), Lab(oratory), Int(erview). Measures: P(ropensity), F(requency), and B(ehaviour). Domains: Alc(ohol), Dri(ving), Dru(gs), Eth(ical), Gam(bling), Gen(eral), Hea(lth)-Gen(eral), Ins(urance), Inv(estment), Occ(upational), Rec(reational), Smok(ing), Soc(ial),

Table S5

Overview and description of the different risk preference measures included in the study, split by measure category, and domain

Category	Domain	Description	Example
Propensity	Driving	Respondents indicate on a (ordinal) scale to what extent they are likely to take risks while driving.	*For each of the following statements, please indicate the likelihood that you would engage in the described activity or behaviour if you were to find yourself in that situation: Not wearing a seat belt when being a passenger in the front seat. Very unlikely/Unlikely/Not sure/Likely/Very likely*
Propensity	Ethical	Respondents indicate on a (ordinal) scale to what extent they are likely to break rules/laws or cause harm to others or the extent to which they identify/perceive themselves as being someone who breaks rules/laws or causes harm to others.	*For each of the following statements, please indicate the likelihood that you would engage in the described activity or behaviour if you were to find yourself in that situation: Taking some questionable deductions on your income tax return. Very unlikely/Unlikely/Not sure/Likely/Very likely*
Propensity	Gambling	Respondents indicate on a (ordinal) scale to what extent they are likely to take risks with gambling-related activities.	*What is the probability that you would do one of the following activities? Please rate on a scale from 0 to 10. Wagering a daily earnings on a bet. 0) very unlikely....10) very likely*
Propensity	Health	Respondents indicate on a (ordinal) scale to what extent they are likely to take risks with regards to their health or take part in activities or make decisions that can have detrimental consequences on their health.	*Please indicate the likelihood that you would engage in the described activity or behaviour if you were to find yourself in that situation: Drinking heavily at a social function. Extremely Unlikely (1) - Extremely Likely (7)*
Propensity	General	Respondents indicate on a (ordinal) scale to what extent they generally identify as someone who likes to take risks or is willing to take risks.	*Are you generally a person who is willing to take risks or do you try to avoid taking risks? Please answer on a scale from 0 to 10, where 0 means "not at all willing to take risks" and 10 means "very willing to take risks".*
Propensity	Investment	Respondents indicate on a (ordinal) scale to what extent they are likely to take risks with investments.	*Which of the statements comes closest to the amount of financial risk that you are willing to take when you save or make investments? Take substantial financial risks expecting to earn substantial returns / Take above average financial risks expecting to earn above average returns / Take average financial risks expecting to earn average returns / Not willing to take any financial risks*
Propensity	Occupation	Respondents indicate on a (ordinal) scale to what extent they are likely to take risks with regards to their job.	*Rate using a scale from 0 to 10. I don't mind taking risks in ... my professional career*
Propensity	Recreation	Respondents indicate on a (ordinal) scale to what extent they are likely to take risks with regards recreational actives or their likelihood of engaging in activities that involve height and/or speed and high risk of serious injury or death.	*For each of the following statements, please indicate your likelihood of engaging in each activity or behaviour: Going down a ski run that is beyond your ability or closed. Very unlikely/Unlikely/Not sure/Likely/Very likely*
Propensity	Social	Respondents indicate on a (ordinal) scale to what extent they are likely to take risks in social situations, or when trusting strangers.	*For each of the following statements, please indicate your likelihood of engaging in each activity or behaviour: Admitting that your tastes are different from those of your friends. Very unlikely/Unlikely/Not sure/Likely/Very likely.*
Frequency	Alcohol	Respondents quantify the extent to which they consumed alcohol or experienced the consequences of alcohol consumption within a certain time frame.	*How many times in the last four weeks have you had an alcoholic drink? Most days / Once or twice a week / 2 or 3 times / Once only / Never*

Table S5 cont.

Measure	Domain	Description	Example
Frequency	Driving	Respondents quantify the extent to which they have not been prudent whilst driving a vehicle within a certain time frame.	*During the past 30 days, how often did you drive a car or other vehicle when you had been drinking alcohol?*
Frequency	Drug	Respondents quantify the extent to which they consumed drugs or experienced the consequences of drug consumption within a certain time frame.	*During the last 30 days, how often, if ever, did you use these other drugs? Heroin, steroids, or MDMA (Ecstasy). 0) Never, 1) Less than once a week, 2) 1 or 2 days per week, 3) 3 or 4 days per week, 4) 5 or 6 days per week, 5) Every day*
Frequency	Ethical	Respondents quantify the extent to which they broke rules/laws or had issues with the law or cause harm to others within a certain time period.	*This is about fare dodging. How often did you do that in the last 12 months? Indicate number of times.*
Frequency	Gambling	Respondents quantify the extent to which they partook in gambling-related activities within a certain time frame.	Pathological gambling (Brodbeck et al., 2009)
Frequency	Occupation	Respondents indicate the extent to which they have been reckless at their job/school or behaved in a way that could lead to them losing their job/get in trouble at school.	*This is about skipping school. How often did you do that in the last 12 months?*
Frequency	Sexual Intercourse	Respondents indicate the number of sexual partners or how often they had sexual intercourse without using a form of contraception within a certain time frame.	*With how many persons are you currently having a romantic or sexual relationship?*
Frequency	Smoking	Respondents quantify the extent to which they smoke cigarettes or other tobacco products within a certain time frame.	*About how many cigarettes or packs do you usually smoke in a day now?*
Behaviour	Gambling	These tasks mention a gambling-related activity/scenario or a form of game. Respondents are asked to decide between two or more options that offer different potential monetary gains and/or losses with varying probability. Also includes Willingness to Pay and Willingness to Accept tasks. Depending on the respondents' responses in such tasks, composite measures can be derived which summarise their tolerance towards risk (e.g., proportion of safe choices, risk aversion category). Such tasks can involve decision from experience or description, with hypothetical or incentivised choices.	*Now, imagine you have a choice between the following two options: Option A: A lottery with a 50% chance of winning 80$ and a 50% chance of losing 50$ / Option B: Zero dollars. Which option would you choose?*
Behaviour	Insurance	Tasks require respondents make choices about insurances, with hypothetical or incentivised choices. Also includes Willingness to Pay and Willingness to Accept tasks. Depending on the respondents' responses in such tasks, composite measures can be derived which summarise their tolerance towards risk (e.g., proportion of safe choices, risk aversion category).	*Assume that you know there is a 50% chance of losing $1000 on a given day. You can take out insurance to cover this amount in case of loss. If an insurance policy is sold as listed below, would you purchase it?*

Table S5 cont.

Measure	Domain	Description	Example
Behaviour	Investment	These tasks explicitly mention an investment-related activity/scenario. Respondents can be asked how much of an endowment they wish to allocate to different options. These tasks can be hypothetical or incentivised.	*Imagine that you had won 100,000 euros in the lottery. Immediately after receiving your winnings you receive the following offer: You have the chance to double your money. But it is equally possible that you will lose half of the amount invested. You can participate by staking all or part of your 100,000 euros on the lottery, or choose not to participate at all. What portion of your lottery winnings would you be prepared to stake on this financially risky yet potentially lucrative lottery investment?*
Behaviour	Occupational	Tasks require respondents to make choices about jobs offering different salaries with different probabilities. Depending on the respondents' responses in such tasks, composite measures can be derived which summarise their tolerance towards risk (e.g., proportion of safe choices, risk aversion category).	*Which ONE do you prefer? Option A: A 50% chance of the salary increasing by 30%, but also a 50% chance of the salary increasing by 11%: Option B: Guaranteed salary increase of 20%.*

End of Table