Universität Basel

Fakultät für Psychologie

Measuring Risk Preference: Temporal Stability, Convergent Validity, and Age Differences

Inaugural Dissertation

In partial fulfilment of the requirements for the degree of Doctor of Philosophy

Submitted to the Faculty of Psychology

University of Basel

by

Alexandra Bagaïni

Born in São Paulo, Brazil

Basel, 2023

Universität Basel

Fakultät für Psychologie

Approved by the Faculty of Psychology at the request of

Prof. Dr. Rui Mata

Prof. Dr. Jörg Rieskamp

Date of the doctoral examination: 31. Aug. 2023

Prof. Dr. Jens Gaab (Dean of the Faculty of Psychology)

Universität Basel

Fakultät für Psychologie

Erklärung zur wissenschaftlichen Lauterkeit

Ich erkläre hiermit, dass die vorliegende Arbeit ohne die Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel selbstständig verfasst habe. Zu Hilfe genommene Quellen sind als solche gekennzeichnet. Die veröffentlichten oder zur Veröffentlichung in Zeitschriften eingereichten Manuskripte wurden in Zusammenarbeit mit den Koautoren erstellt und von keinem der Beteiligten an anderer Stelle publiziert, zur Publikation eingereicht, oder einer anderen Prüfungsbehörde als Qualifikationsarbeit vorgelegt. Es handelt sich dabei um folgende Manuskripte:

1. Bagaïni, A., Liu, Y., Kapoor, M., Son, G., Bürkner, P.C., Tisdall, L., & Mata, R. (2023). Comparing the Temporal Stability and Convergent Validity of Risk Preference Measures: A Meta-Analytic Approach. *Manuscript Submitted for Publication.*

2. Bagaïni, A., Liu, Y., Bajrami, A., Son, G., Tisdall, L., & Mata, R. (2023). Aging and Economic Preferences: Cumulative meta-analyses of age differences in risk, time, social, and effort preferences. *The Journals of Gerontology: Series B, 78*(8), 1122-1135. doi: 10.1093/geronb/gbad034

3. Liu, Y., Bagaïni, A., Son, G., Kapoor, M., & Mata, R. (2023). Life-course trajectories of risk-taking propensity: A coordinated analysis of longitudinal studies. *The Journals of Gerontology: Series B, 78*(3), 445-455. doi:10.1093/geronb/gbac175

Universität Basel

Fakultät für Psychologie

Spezifizierung des eigenen Forschungsbeitrags zu den Manuskripten:

1. Eigener Beitrag nach CRediT[1]:

☒ Conceptualization	☒ Data curation	☒ Formal Analysis
☐ Funding acquisition	☒ Investigation	☒ Methodology
☒ Project administration	☐ Resources	☐ Software
☐ Supervision	☐ Validation	☒ Visualization
☒ Writing – original draft		
☒ Writing – review & editing		

2. Eigener Beitrag nach CRediT[1]:

☒ Conceptualization	☒ Data curation	☒ Formal Analysis
☐ Funding acquisition	☒ Investigation	☒ Methodology
☒ Project administration	☐ Resources	☒ Software
☐ Supervision	☐ Validation	☒ Visualization
☒ Writing – original draft		
☒ Writing – review & editing		

3. Eigener Beitrag nach CRediT[1]:

☐ Conceptualization	☐ Data curation	☐ Formal Analysis
☐ Funding acquisition	☒ Investigation	☐ Methodology
☒ Project administration	☐ Resources	☐ Software
☐ Supervision	☐ Validation	☐ Visualization
☐ Writing – original draft		
☒ Writing – review & editing		

[1] https://casrai.org/credit/

Universität Basel

Fakultät für Psychologie

Open-Science Aspekte der Manuskripte:

1. Preregistration: ☐ ja ☒ nein
   Open-Access-Publikation: ☒ ja ☐ nein
   Open-Access-Data/Analyse: ☒ ja ☐ nein
   Ort/URL der Daten und Analysen: https://osf.io/5kzgd/

2. Preregistration: ☐ ja ☒ nein
   Open-Access-Publikation: ☒ ja ☐ nein
   Open-Access-Data/Analyse: ☒ ja ☐ nein
   Ort/URL der Daten und Analysen: https://github.com/cdsbasel/cumulative

3. Preregistration: ☐ ja ☒ nein
   Open-Access-Publikation: ☒ ja ☐ nein
   Open-Access-Data/Analyse: ☒ ja ☐ nein
   Ort/URL der Daten und Analysen: https://github.com/cdsbasel/ageriskmeta

Ort, Datum Basel, 11/07/2023

Vorname Nachname Alexandra Bagaïni

Acknowledgements

First, I am very grateful to my supervisor, Prof. Dr. Rui Mata, for all his encouragements, patience, and the time he invested in my work. The many interesting and dynamic discussions we had as well as his feedback were very enriching and greatly helped improve my work. I'd like to also thank Prof. Dr. Jörg Rieskamp for sharing his time and advice, and for also having the opportunity work with him in the context of an interdisciplinary project.

I would like to thank Dr. Loreen Tisdall, for being such a supportive colleague, skilled collaborator, and for taking the time to provide helpful feedback and advice. Thank you to Yunrui Liu, with whom I shared my PhD journey and had the opportunity to collaborate on projects. I would also like to thank Laura Willes for correcting manuscripts, and helping with administrative tasks.

Thank you to the rest of the Cognitive and Decision Sciences team, for their help and assistance in the different projects, and who I had the chance to meet and interact with. A special thanks to current and previous research assistants, Gayoung Son, Madlaina Kapoor, and Arzie Bajrami, for screening all these abstracts and going through codebooks.

Thank you to all my friends and family for providing continuous motivation and support throughout the process of writing this dissertation.

Finally, I am very grateful to my parents, Christine and Michel, for believing in me and consistently supporting me during the ups and downs of doing a PhD.

Table of Contents

Abstract

Across various domains in life we need to make choices where the outcomes are not guaranteed and there is the potential for a loss. Individuals differ in their willingness to partake in risky activities or make choices under risk. Risk preference, is a psychological construct that reflects individual's appetite for risk. Various disciplines, in particular, psychology and economics, have developed risk preference measures and used these to investigate inter and intra-individual differences. Despite the popularity of risk preference in the behavioural sciences, we lack a clear understanding of how stable this construct is and how coherently it is captured. This lack of clarity can have consequences on how well we understand and quantify individual differences in risk preference, in particular age differences. This dissertation aims to address these open questions by using meta-analytic methods, where we synthesised and analysed data from various sources. In three studies we: (1) compare the temporal stability and convergent validity of risk preference measures; (2) assess to what extent published evidence on age differences in task-based risk-taking aligns with theoretical predictions; and (3) how self-reported risk-taking propensity changes across adulthood. Overall, (1) we observe substantial differences in the temporal stability of risk preference measures and an overall lack of convergence; (2) whilst most theories predict an age-related decline in risk taking, this is not in line with the evidence observed from behavioural tasks; in contrast (3) we note across several domains, that self-reported risk-taking propensity declines with age. Through these three studies, we show that not all measures of risk preference are comparable, and that we need to establish a clearer definition and operationalisation of the construct. This has implications for the understanding of individual differences, as well as the development and evaluation of theories.

Introduction

Risk is prevalent in many aspects of an individual's daily life, such as the type of hobby that they practice, the job they do, or how they manage their money. Individuals partake in such activities and make choices in these different situations even though it is not guaranteed to what extent they will experience a win or a loss. The way individuals navigate risk can have an important impact on their well-being. Risk preference, a psychological construct that is generally defined as an individual's appetite for risk, can impact the decisions that are made across various life domains. For instance, in the occupational domain, studies have found that individuals who are more risk-tolerant, are more likely to become self-employed (Beauchamp et al., 2017). In the financial domain, we observe that more risk-tolerant individuals invest more often in stocks (Dohmen et al., 2011). The importance of an individual's risk preference is also illustrated by the requirement of financial institutions to establish a risk profile of their client prior to assigning them a specific product (Financial Services Authority, 2011). Furthermore, in the health domain, risk preference has been linked to tobacco and alcohol consumption (Beauchamp et al., 2017; Yang et al., 2022), as well as the probability of having a health insurance (Kagaigai & Grepperud, 2023).

Across different disciplines, notably in psychology and economics, definitions and measures of risk preference differ. In psychology, risk preference is assessed by the propensity to engage in behaviours or activities that whilst rewarding also carry a chance for loss or injury, on the other hand, in economics, risk preference is linked to favouring varying monetary payoffs over (more) certain ones (Hertwig et al., 2019; Mata et al., 2018). From these definitions, three measure categories emerge:

1. Propensity measures are (direct) self-reports of respondent's liking for risk or willingness to take risks in general or in specific domains (e.g., *Rate the following statement: I like risk. 1(Not like me at all) – 4 (Very much like me)*);

2. Frequency measures are self-reports of the rate at which respondents take part in certain domain-specific risky activities (e.g., *On average, how many units of alcohol do you consume in a week?*).
3. Behavioural measures are tasks involving (monetary) payoffs and different probabilities, such as gambles (e.g., 50-50 chance of winning $20 or losing $5). Based on the respondents' choices, their utilities are determined. These measures are more consistent with economics' definition of risk preference.

Taking into consideration these different definitions and operationalisations of risk preference, it raises the question of how coherently this construct is being evaluated, and the possible impact this can have for how accurately individual differences, namely age differences in the context of this dissertation, are captured. Further, to adequately measure such differences and consequently predict behaviour, the measures used must not only be reliable, but must also capture a set of behaviours or attitudes, that are not overly situation-specific or prone to change over time (e.g., Anusic et al., 2012; Enkavi et al., 2019). In the following sections, I expand on the relevance for risk preference to accurately quantify (1) temporal stability, (2) convergent validity, and (3) age differences, as well as how this dissertation will address the current gaps in the literature.

Temporal Stability

To understand the development of cognitive functions, or establish the long-term effects of certain life experiences, or assess the efficacy of interventions, longitudinal studies are key. In these studies, biological, behavioural, and/or survey data is collected multiple times across a certain time period from the same set of individuals, which can result in a very rich and informative set of data (e.g., The Dunedin Study).

In psychology, a substantial amount of research is conducted to understand the stability and change of psychological constructs (e.g., Bleidorn et al., 2022; Mund et al., 2020; Orth,

2018). Temporal stability can be quantified in two ways: (1) rank-order stability, which is how consistent the rank ordering of individuals is over time; and (2) mean-level stability, which refers to how consistent the average level of a characteristic is over time (Josef et al., 2016).

With the start of longitudinal studies, researchers were quickly interested in ranking the consistency of different psychological constructs, which introduced a continuum (or hierarchy) of consistency (Conley, 1984; Crook, 1941; Darley, 1938). In this continuum, on one end there would be traits, which are attributes or characteristics that are enduring and less permeable to changes in the environment, and on the other, there would be states, which are characteristics that are short-lived and responsive to changes in the environment (Anusic et al., 2012; Conley, 1984). In this field of study, it was quickly observed that intelligence was one of the most consistent constructs, followed by personality traits, and then social attitudes (cf. Conley, 1984). As it has been argued that there are no psychological constructs that are fully stable or constantly changing, it is important to take advantage of the development of quantitative methods to better capture where along the continuum different constructs lie (Anusic & Schimmack, 2016).

Having an accurate description of the extent and when a certain characteristic or set of behaviours are more or less permeable to change offers valuable insights. First, it allows to better plan and implement effective interventions. For example, by knowing that certain behaviours are less stable at adolescence than at adulthood, it is coherent to design an intervention targeted at adolescents, as they would likely be more receptive and affected by the treatment (Anusic et al., 2012; Conner & Norman, 2022). Second, and relatedly, it opens opportunities for research to investigate the factors associated with these changes. Third, it can help improve the prediction of behaviour or certain life outcomes, as stable psychological traits in comparison to situation-specific psychological states, are more useful to make meaningful predictions (e.g., Stachl et al., 2020).

Whilst considerable research was conducted on the rank-order stability of psychological constructs such as personality, intelligence, well-being and life-satisfaction (e.g., Anusic & Schimmack, 2016; Deary, 2014), comparatively less attention was placed on risk preference despite its relevance in various research fields and real-world applications. A few exceptions, include the work by Josef et al. (2016) who used data from the German Socioeconomic Panel, which included responses spanning a period of 10 years. They reported for both domain-general and domain-specific (e.g. driving) risk-taking propensity, rank-order stability estimates ranging between 0.4 and 0.5, and akin to personality traits, these estimates followed an inverted U-shape trend with age (Josef et al., 2016). Suggesting more change in risk-taking propensity in younger and older adults. Frey et al. (2017) used a wider set of measures, and found that 6-month test-retest correlations varied as a function of measure category, with higher test-retest correlations for propensity and frequency measures than for behavioural measures.

On the basis of the available evidence, however, we cannot adequately establish the temporal stability of risk preference. To do so, we first need to properly disentangle measurement reliability from *real* change in the construct, and thus analysing data collected at both short and long time intervals is pivotal (Anusic & Schimmack, 2016). And second, given the multiple definitions of risk preference in the literature, it is important to account for differences between measure categories (i.e., propensity, frequency, and behaviour) and domains (e.g., health, driving) when assessing its stability, which is currently lacking. In Manuscript 1 we specifically address this gap, by using longitudinal panel data to conduct an individual-participant data meta-analysis and quantify the temporal stability of risk preference as well as assess the effects of measure category and domain.

Convergent Validity

How psychological constructs are defined and measured has implications for how predictions are derived from theory, how these theories are developed and evaluated as well as how research findings are replicable and comparable (Bringmann et al., 2022; Protzko et al., 2020; Strickland & Johnson, 2021). Generally, and as already shown for risk preference, various measures can be used to assess a *single* psychological construct, for instance, a survey identified 280 different measures used to assess depression (Santor et al., 2006, as cited in Fried et al., 2022), and a meta-analysis on self-control found 100 unique self- and other-report questionnaires (Duckworth & Kern, 2011). The assumption that these different measures are all targeting the same construct, is not generally supported by empirical evidence. For instance, the convergence between different self-control measures ranged between approximately 0 and 0.35 (Duckworth & Kern, 2011). Even lower estimates (r = 0-0.15) were reported for measures of empathy (Murphy & Lilienfeld, 2019). This warrants caution when deriving conclusions from results that stem from a single measure.

Issues associated with the development and application of different measures, partly stem from a lack of agreement on the definition of constructs (Bringmann et al., 2022). Further, this lack of clarity is also prevalent in theories, which are generally agnostic to the mode of operationalisation, this has an impact on the precision of the hypotheses that are derived from them, and the ability to falsify them (Eronen & Bringmann, 2021). Relatedly, this questions the adequacy of translating evidence from one form of measurement to another (Strickland & Johnson, 2021).

A manner to better understand the lack of convergence between measures, is by differentiating between behavioural and self-report measures. Behavioural measures, are designed to increase within-person variance in response to a particular treatment or manipulation, with little capacity to capture individual differences in the treatment effect

(Dang et al., 2020). On the other hand, self-report measures do not share these features, and are better able to capture between-person variability resulting in greater reliability (Hedge et al., 2018). This results in what has been named a reliability paradox whereby, measures that do not capture between-subject variability cannot be highly correlated with other constructs, which limits to what extent these measures can be used to study or predict individual differences (Enkavi et al., 2019; Hedge et al., 2018).

Furthermore, by design, these measures tend to create different contexts and thus solicit different response processes (Dang et al., 2020). Behavioural measures, create a very structured context with a clear set of stimuli and instructions, however, self-report measures are not as structured, as they inquire individuals to reflect on their everyday life. As a result, different response modes are at play, which further minimize the associations between these measures despite being aimed at assessing the same construct (Dang et al., 2020).

To derive accurate conclusions from the associations between measures, these must be based on the responses from a large enough set of participants or observations. A simulation study found that to compute *stable* correlations, a sample of at least 250 individuals is advised (Schönbrodt & Perugini, 2013). This can be of particular concern for task-based functional neuroimaging studies, that ofttimes have sample sizes below 50 (Elliott et al., 2020). In the recent years, as functional neuroimaging studies have become more accessible, a growing number of individual differences research has begun to include biological data or biomarkers (e.g., region-specific brain activity) as variables of interest. However, in the last few years, concerns have been raised regarding the suitability of certain biomarkers for this type of research (Gratton et al., 2022; Marek et al., 2022). In the context of a meta-analysis and re-analysis of two datasets, Elliot et al., (2020) reported low intra-class-correlation coefficients for the reliability of task-fMRI measures (e.g., N-back memory task, face recognition). Additionally, a study using the UK Biobank, reported that data from 1,500 - 3,900 participants

would be needed to produce reliable associations between functional or structural brain measures with different phenotypes, such as intelligence, or alcohol consumption (Liu et al., 2023). Overall such studies highlight the importance of measure reliability to adequately understand and predict individual differences.

As described above, there are various ways to assess risk preference, and studies that have compared commonly used risk preference measures with each other, found that correlations between them are relatively low (Frey et al., 2017; Pedroni et al., 2017). Unlike other constructs (e.g., self-control: Duckworth & Kern, 2011; Sharma et al., 2014) a comprehensive analysis of the convergence between risk preference measures is lacking. Such a summary can help determine how cohesive the picture of risk preference is. In Manuscript 1 we additionally make use of longitudinal panel data to (1) assess the correlations between a wide range of risk preference measures in fairly large samples; and (2) use estimates of reliability stemming from the analysis of temporal stability to further understand the presence or lack of convergence between measures.

Age Differences

By 2050, individuals aged 65 are predicted to live an additional 19 years versus 17 years in 2020 (United Nations, 2019). With an ageing global population, and the relevance of risk preference in different life domains, particularly for financial decisions, understanding to what extent risk preference differs between age groups, and how it changes across the lifespan can have important societal implications (Schildberg-Hörisch, 2018).

From infancy to late adulthood, we experience changes at the biological, cognitive, and socio-economic levels, which impact how we make decisions. Further, as we transition in different phases of our lives, the relevance of certain decisions change (i.e., saving for retirement, changing careers). Age has been a factor that numerous studies have examined to better understand how decision-making changes (e.g. Frey et al., 2021; Seaman et al., 2022;

Sparrow et al., 2021; Westbrook et al., 2013). Similarly, theories posit that age is a key variable that contributes to changes in individual's behaviour, such as risk-taking. For instance, as presented by the dopaminergic neuromodulation hypothesis, a decline in dopaminergic functioning results in a reduction of older adult's responses towards rewards, thus making it less appealing take risks to obtain larger ones (cf. Frey et al., 2021). At the socio-economic level, the risk-sensitivity hypothesis suggests that as individuals become older, their financial capital grows and social network increase in size, and in turn this results in a reduced need to take risks (cf. Frey et al., 2021).

A meta-analysis of risk preference as measured by risk-taking in behavioral tasks, found that age differences between young and older adults depended on the type of task or domain (Best & Charness, 2015; Mata et al., 2011). Yet, Josef et al. (2016) found a quadratic relation between age and risk taking, a trend also observed for self-reported measures, even though these were not highly correlated with each other. Assessments of the robustness of these age-related effects in risk taking is currently lacking. Such information is important to understand the current status and the potential need for additional evidence on age differences in risk taking, as well as its generality across populations. In Manuscripts 2 and 3 we aim to address these gaps by using meta-analytic methods to (1) assess the magnitude and robustness of age differences on risk preference and other economic preferences reported in the literature; and (2) compare the magnitude of age differences in self-reported risk-taking propensity across multiple longitudinal data sets and domains, respectively.

Overview of Manuscripts

In this section, I describe how the work conducted across three studies contribute to the understanding of the measurement of risk preference in general, and in the context of research on age differences. This work aims to provide a comprehensive overview of the convergence and temporal stability of risk preference measures, as well as an assessment of the extent that

age differences are observed across different measure categories and domains. Specifically, in Manuscript 1, using a meta-analytic approach and multiple sets of longitudinal data, we explore the rank-order stability and convergence of risk preference measures by considering effects of measure category, domain, and age. In Manuscript 2, we focus on economic preferences, including risk preference, as measured by behavioural tasks, and conduct a meta-analysis to evaluate the robustness of the published evidence on age differences and its consistency with theoretical expectations. Lastly, in Manuscript 3, we examine changes in risk-taking propensity across the life span.

Manuscript 1: The Temporal Stability and Convergent Validity of Risk Preference Measures

The work presented in Manuscript 1 daws on the analytical approach presented in Anusic & Schimmack (2016) and extends on the work of Frey et al. (2017). The meta-analytic model of stability and change (MASC) introduced by Anusic & Schimmack (2016) aims to capture the trajectory of test-retest correlations over time by distinguishing between measurement error and true change. It is a non-linear model that includes three parameters: reliability, change and stability of change. Reliability represents the true proportion of between-person variance, change represents the proportion of that reliable variance that is prone to change, and stability of change is the rate at which change happens over time. To test this model, the authors collated a set of test-retest correlations with retest intervals of up to 15 years from four psychological constructs: personality, self-esteem, life satisfaction, and affect. In comparison to other constructs, they observed that personality was the most reliably measured construct as well as the most stable, whilst, affect was the least (Anusic & Schimmack, 2016).

Frey et al. (2017) collected on two occasions data using an extensive battery of risk preference measures comprised of propensity, frequency, and behavioural measures. Yet, with

a retest interval of 6 months and a sample size of just over 100 participants, it is not sufficient to properly estimate robust and representative reliability and stability estimates. In Manuscript 1, using data from longitudinal samples, we address the open questions of a) where along the trait-state continuum does risk preference lie, and b) to what extent different measures of risk preference are correlated.

For this study, we selected longitudinal samples that had data on risk preference measures across at least two time points. Using a number of criteria, we selected measures that spanned across three categories (i.e., propensity, frequency, and behaviour), from various domains (e.g., smoking, alcohol). For each measure, we coded key information, such as category, domain, and type of response scale. For each longitudinal sample, and each measure, we computed test-retest correlations for every possible combination of waves, and did so separately for male and female respondents of different age groups. Furthermore, for samples that contained at least two measures of risk preference, we calculated the inter-correlations between measures (only between responses collected at the same data collection point). By following this approach, we obtained over 72,000 test-retest correlations with test-retest intervals ranging from a couple of weeks to 20 years, and over 60,000 inter-correlations. This included data from over 500,000 unique individuals, and over 300 measures of risk preference.

Using the set of test-retest correlations, we conducted two analyses: variance decomposition (Grömping, 2007), and a meta-analysis using MASC (Anusic & Schimmack, 2016). In both analyses we were interested in the effects of respondent (e.g., age) and measure-related (e.g., domain) variables. In the variance decomposition analyses, we found that domain explained substantially more variance in the test-retest correlations of frequency measures (12.5%) than that of propensity (1.3%) or behavioural measures (5.6%). Age explained less than 1% for behavioural measures but explained 8.4% for frequency measures. Lastly, retest interval explained 5.2% and 6.9% of the variance for propensity and frequency measures,

respectively, and only l% for behavioural measures. Such results already suggest differences between measure categories, and serve as a good rationale for accounting for the effects of age, and domain on the temporal stability of risk preference. Specifically, when estimating the MASC model for each set of test-retest correlations (i.e., separately for propensity, frequency and behaviour), we were interested in the effects of age, domain and gender on all three parameters.

For reliability, we observed overall clear differences between measure categories which echo that of previous research (Frey et al., 2017). Specifically, propensity and frequency measures were on average more reliable than behavioural measures. In addition, there were substantial domain differences in the reliability of frequency measures, with smoking being the most reliable, and ethical actions the least. We did not observe such prominent domain differences for propensity or behavioural measures. Regarding age trends, similar to the work by Josef et al. (2016), we observe for both propensity and frequency measures an inverted U-shape, however we do not observe such trends for behavioural measures.

For the level of change and the rate of change that we observe over time, we noted that both tobacco and alcohol consumption were relatively stable, meaning that once an individual starts to smoke or consume alcohol, their consumption is not going to drastically change. In contrast, we noted that acts of violence or breaking the law (i.e., ethical domain) were more prone to change. We observed less drastic domain differences for propensity and behaviours. We additionally re-analysed the data from the study conducted by Anusic & Schimmack (2016), in comparison to personality traits, propensity and frequency measures were less reliable, and overlapped with estimates of affect.

Using the set of inter-correlations, we conducted a variance decomposition analysis and a Bayesian meta-analysis. From the variance decomposition analysis, we observed that more than half of the explained variance between inter-correlations could be explained by whether

or not the domains and category of both measures were the same (17.5% out of 26.6% of explained variance). Furthermore, we found that unlike test-retest correlations, age was not a key predictor. Lastly, unlike it was proposed, we found no support for the idea that lack of reliability could explain low convergence (e.g., Dang et al., 2020).

Overall, from the meta-analysis, we observed that risk preference measures were poorly correlated (M = 0.16), but this overall estimate concealed substantial heterogeneity. Specifically, at the category-level, convergence within measure categories (0.19-0.41) was greater than that between categories (0.02-0.14). At the domain-level, meta-analytic estimates ranged between -0.2 and 0.8, and pairs of measures with the same (versus different) domain and category had overall higher meta-analytic estimates. Taken altogether, these measures do not currently paint a unified picture of risk preference.

In conclusion, across measures, we observed diverging trends in temporal stability, and overall low inter-correlations. Such results question to what extent we are capturing a single construct, and suggest that risk preference currently lacks conceptual clarity.

Manuscript 2: Age Differences in Economic Preferences

To choose amongst a set of options, individuals weight and compare the benefits and costs. Depending on the choice context (e.g., investment, donation, savings), these decisions involve different forms of benefits and costs, and are therefore guided by different types of preferences. Economic preferences reflect the trade-offs individuals make between monetary benefits and costs such as risk, time, selfishness and effort (Soutschek & Tobler, 2018). These preferences, are commonly measured using behavioural tasks in which the monetary outcomes (e.g., $5 versus $10) and probabilities (e.g., 50% and 25%), or waiting time (today versus 10 days), or effort level (20% versus 70% of maximum strength) or closeness with others (e.g. neighbour versus friend) are manipulated, and an index (e.g., discounting rate, proportion of choices) is computed based on the choices made.

Given that economic preferences have an important impact on our everyday decisions, there has been considerable research conducted and theories developed to further understand how and why individuals differ in their economic preferences (e.g., Frey et al., 2021). In this study we focused on age differences in risk, time, social and effort-related preferences. By conducting a survey of theories, we noted that in general these predict a decline in risk taking and temporal discounting with age; whilst predicting an increase in effort discounting and altruism with age.

Meta-analyses conducted on age differences in risk preference (Best & Charness, 2015; Mata et al., 2011), time discounting (Seaman et al., 2022), and altruism (Sparrow et al., 2021), have yielded mixed results regarding the strength of evidence for the existence of age differences, and their consistency with related theories. Yet, thus far the effect of age on these preferences have been meta-analysed separately, using different methods and criteria. Given their relevance and inter-relatedness in everyday decisions, it can be insightful to assess them altogether under comparable conditions. In Manuscript 2, we conducted a synthesis of the literature on age differences in economic preferences, investigated the robustness of this evidence, and assessed the degree to which theoretical predictions matched the empirical evidence.

We first updated previous meta-analyses (i.e., Best & Charness, 2015; Mata et al., 2011; Seaman et al., 2022; Sparrow et al., 2021), and conducted a new search for studies on age differences in effort discounting. For each economic preference, we computed an overall meta-analytic estimate, and conducted a set of meta-regressions to further understand the heterogeneity in the effect of age across studies. With our approach we were able to include moderators that were common across all four preferences (e.g., incentivization, study design), as well as preference-specific moderators (e.g., gain/loss/mixed domain for risk preference).

In addition, we conducted cumulative meta-analyses (Lau et al., 1992) to account for how evidence accumulated over the years and its robustness. A cumulative meta-analysis consists of repeatedly estimating meta-analytic estimates by gradually integrating the evidence of new studies. It is an approach that allows to assess how stable the evidence is, and examine any effect of publication bias (Clarke et al., 2014; Hopewell et al., 2005; Mullen et al., 2001). In such a way we can better assess the strength of evidence linking economic preferences and age.

Overall, we observed small effects of age across all four economic preferences. In our analyses we identified non-significant effects of age for risk (r = -0.02, 95% CI[-0.06, 0.02]), and effort (r = 0.24, 95% CI[-0.05, 0.52]) preferences, and a small but significant effect of age for social (r = 0.11, 95% CI[0.01, 0.21]) and time (r = -0.04, 95% CI[-0.07, -0.01]) preferences. These results suggest more altruism and patience with age. However, when accounting for equivalence tests, these effects were not significantly distinguishable from an equivalence bound of |.1|.

The cumulative meta-analyses revealed that for both risk and time preference, very early on, meta-analytic estimates moved close to zero and did not substantially change over time, questioning to what extent additional evidence on age differences is required for risk and time preferences. Interestingly, for time preference, we found evidence of the Proteus Phenomenon (Ioannidis & Trikalinos, 2005; Young et al., 2008), whereby the large difference between older and young adults reported in the first published paper (i.e., Green et al., 1994) was not replicated by subsequent studies.

Regarding the effect of moderators, we found a negative effect of age in the gain domain for risk preference, which is consistent with past work (Best & Charness, 2015). Contrary to the meta-analysis by Mata et al. (2011) we found no effect of task type (i.e., experience versus description-based decision-making) on age differences.

In general, such small effect sizes provide weak evidence of age differences, and are not in line with theoretical expectations. We additionally observed, that unlike risk and time preferences, less research has been conducted on age differences in social and effort-related preferences and relatedly, we did not observe *stable* estimates, which calls for the analysis of additional evidence. Furthermore, it is important to acknowledge that the findings presented in this study might also depend on the measures used to quantify these economic preferences. In this meta-analysis, with the aim of preserving comparability between preferences, we adopted a restricted definition by focusing on a specific outcome (i.e., monetary) and operationalisation (i.e., behavioural tasks). The theories described in the context of this study were however vague regarding the definitions and operationalisation of the constructs. In Manuscript 3, we focused on risk preference again, and used self-report measures to assess, if akin to previous work (Dohmen et al., 2017; Josef et al., 2016), there were observable age-related changes in risk-taking.

Manuscript 3: Life-Course Trajectories of Risk-Taking Propensity

As introduced in Manuscript 2, theories on development generally posit that risk-taking decreases with age (e.g., signalling hypothesis, cf. Frey et al., 2021). However, in Manuscript 2 we find no strong evidence of an age effect in the published literature. In that meta-analysis we focused on behavioural measures of risk preference, and as shown in Manuscript 1 and other related work (Frey et al., 2017), these were the least reliable measures, which question their suitability to capture meaningful individual differences (Dang et al., 2020; Enkavi et al., 2019). In contrast, measures of risk-taking propensity were more reliable, and therefore in Manuscript 3 we focus on these to study age differences.

Despite a number of cross-sectional and longitudinal work examining the effect of age on risk-taking propensity (e.g., Dohmen et al., 2017; Josef et al., 2016; Mata et al., 2016), a robust quantitative assessment of its trajectory across adulthood is currently lacking. Such a

comprehensive overview can contribute to assessing the replicability of these differences, and in turn the applicability of theories of adult development to explain changes in risk-taking behaviour across the lifespan. With this objective, we used a coordinated analysis to investigate age trends in mean-level change in risk-taking propensity across various data sets and domains.

We first compiled a list of longitudinal samples that included measures of self-reported risk-taking propensity in seven domains (e.g., general, financial, social). For each sample and domain, we first estimated different multilevel models that either included or excluded a series of effects (e.g., linear versus quadratic age effect), and via model comparison, we selected the best-fitting model. In a second step, we meta-analysed the estimates for each domain.

Based on the estimates of the best fitting model, across all domains and samples, we detected a negative (linear) effect of age, and this decline was steeper for certain domains (recreational) than others (health). As we also included gender in the model, we noted that across all domains, males indicated higher risk-taking propensity than females. However, there was considerable heterogeneity between the estimates of age effects, and a variance decomposition analysis revealed that both domain and sample accounted for a substantial amount of this variance. Lastly, the results of the meta-analyses showed that the meta-analytic effect of age was generally less pronounced than that of gender.

Before concluding, we note that in this work we focused on synthesizing and describing the available evidence, and unlike related work (e.g., Malmendier & Nagel, 2011), because of a lack of comparable variables across samples, we did not assess the impact that certain life events can have on individuals' risk-taking propensity, and thus provide more concrete reasons for why we observe these trajectories across the lifespan.

Overall, the age-related decline that we observe in risk-taking propensity is compatible with several theoretical accounts (cf. Frey et al., 2021). Furthermore, the relevance of domains to explain variance between estimates highlights the need for theory to distinguish between

these, so as to provide more detailed accounts of the mechanisms underlying age-related differences. Lastly, based on our results, we note that to make predictions on risk attitudes, accounting for domain and population-related factors can have a potential benefit.

Discussion

In this dissertation I aimed to provide a more comprehensive overview of the convergence and temporal stability of risk preference measures, and to assess how age differences are captured. From this work, three main conclusions can be derived. First, there is considerable heterogeneity in the reliability and temporal stability of risk preference measures. Second, we observe a low convergence between these measures, thus drawing a rather disunited picture of risk preference. Third, given these discrepancies, we did not observe age differences across all measure categories or domains, thus not consistently supporting the general theoretical expectations of an age-related decline in risk taking.

By addressing current gaps in the literature using meta-analytic methods, the work presented in this dissertation contributes to the research on risk preference by raising concerns about its measurement, and evaluating the suitability of certain measures given their properties (e.g., reliability) to accurately investigate age differences. In the sections that follow, I discuss the implications of this overall work, and avenues for future research.

Implications

The findings presented in this dissertation have implications for (1) the conceptualisation of risk preference; (2) the discussion and interpretation of results from individual differences research on risk preference; and (3) showcasing the use of available (longitudinal) data sets for research.

Concerns have already been raised about the lack of conceptual clarity of psychological constructs in general, and its impact for measurement and theory development (Bringmann et al., 2022). The evidence presented in this dissertation show concretely how this applies to risk

preference. In particular, theories of development are currently too vague and do not explicitly integrate measurement or domain information in its predictions or the descriptions of the mechanisms underlying age differences in risk preference. Therefore, these theories, by their vagueness, cannot be properly challenged or falsified (Meehl, 1990). Importantly, the evidence presented in this dissertation cast doubts on the robustness of age effects in risk preference, and hence its suitability as a phenomenon to establish good theories (Eronen & Bringmann, 2021).

Moving forward, these insights should encourage individual differences research on risk taking to be more sensible about the quality of the measures used, the robustness of potential associations between variables, and the extent that generalisations are adequate, as these can be important for replication (Protzko et al., 2020). In particular, there needs to be an awareness that certain measures (e.g., risky choices in a gamble task and self-reported risk taking-propensity) should not be used interchangeably, and thus that certain results, might be measure-specific. Based on the evidence we gathered on temporal stability, results might also be *time*-specific or not easily replicable as measures are either not reliable or do not capture behaviours that are stable over time. It is rare that researchers address the stability of the association between variables in the interpretation of their results. By quantifying the temporal stability of a construct, comparing it to others, and placing it on a trait-state scale, can potentially encourage this discussion.

It can be a challenging endeavour to collect enough data to reasonably claim that the observed results in the context of a study are robust and generalizable. It is not always possible or sensible (e.g., due to budget constraints) to collect data using a battery of more than a handful of measures or more than once. Taking this aspect into account, the advantages of using household survey data become more evident. Increasingly such surveys integrate more experiment-like measures (e.g., Understanding Society-Innovation Panel) and collect responses from a relatively large number of respondents (i.e., $n > 1{,}000$) that are presentative

of the population. In this dissertation we show how this is possible, and the value of meta-analysing such data to obtain robust estimates of temporal stability, convergent validity, and age differences. For Manuscripts 1 and 3, we were able to compile a rich collection of risk preference measures, and synthesise the responses from a large number of participants. It would have otherwise been very costly to collect such an amount of data on our own. With Open Science becoming more prominent, accessing raw data from published research is easier, and where possible we integrated this in the work described in Manuscript 2.

Future Directions

In the current work we focused on behavioural and self-report measures to understand risk preference as a construct, and to explore age differences. However, an element currently lacking in this measurement *equation,* are objective measures, such as financial or medical records, as well as sensor data. To establish appropriately the external validity of measures at hand such data is vital. Therefore assessing the extent that responses from objective, behavioural and self-report measures overlap, can in addition to reliability, help assess the quality of these measures to capture real-world behaviour (e.g., digital-media use: Parry et al., 2021), and their relevance for prediction (e.g., criminal behaviour: Epper et al., 2022).

As different individuals can have different definitions of what it means to take risks (Arslan et al., 2020; Steiner et al., 2021), future research could take advantage of the growth and availability of open-source language models and text-based data to capture individual differences (Wulff & Mata, 2022) and potentially use it to better predict and understand different aspects of risk preference.

Here, we analysed data from observational studies, but experimental and quasi-experimental data are pivotal to directly assess the underlying mechanisms presented by theories. Across the three projects, given the data that was analysed, we could not directly make claims on the factors that cause risk preference to be less stable at young adulthood, or about

the mechanisms underlying the relationship between age and risk preference, or quantify the impact of a certain life events (e.g., marriage, unemployment). Future work could implement such analyses, by factoring in the analyses additional variables, such as the occurrence of political, societal or environmental events (e.g., Malmendier & Nagel, 2011).

Lastly, whilst we used estimates of temporal stability to compare risk preference to other psychological constructs and place it along a trait-state continuum, we could not do the same for convergent validity. Research on individual differences would greatly benefit from having comprehensive overviews of the convergence of measures of psychological constructs, especially those with an extensive measurement history (e.g., depression). Such work has been conducted for self-control, empathy, as well as emotional intelligence (cf. Dang et al., 2020). With the assessment of more constructs, potentially a "convergence-divergence" continuum can be proposed: on one end, psychological constructs with measures that are highly correlated with one another, thus creating a unified image of the construct; and on the other end, psychological constructs with measures that are very poorly correlated, thus drawing a less cohesive image. Conducting such syntheses can provide an overview of available measures, how much these can be used interchangeably, and thus critically evaluate the conceptual clarity of different psychological constructs. Thus, placing risk preference within a clearer context with regards to the convergence of its measures.

Conclusion

Risk is a key element of everyday decision-making, and risk preference shapes how individuals approach it. In this dissertation I looked at how risk preference is measured. Currently, the measures used do not speak with one voice, and in a way, each tell a different story about an individual. To effectively capture the meaningful stories, moving forward, we should acknowledge this measurement heterogeneity, and become more mindful of the measures that we use, so as to design more replicable studies and use resources more wisely.

References

Anusic, I., Lucas, R. E., & Donnellan, M. B. (2012). Dependability of personality, life satisfaction, and affect in short-term longitudinal data. *Journal of Personality*, *80*(1), 33–58. https://doi.org/10.1111/j.1467-6494.2011.00714.x

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), Article 1. https://doi.org/10.1038/s41598-020-72077-5

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

Best, R., & Charness, N. (2015). Age differences in the effect of framing on risky choice: A meta-analysis. *Psychology and Aging*, *30*(3), 688–698. https://doi.org/10.1037/a0039447

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

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

Clarke, M., Brice, A., & Chalmers, I. (2014). Accumulating Research: A Systematic Account of How Cumulative Meta-Analyses Would Have Provided Knowledge, Improved

Health, Reduced Harm and Saved Resources. *PLOS ONE*, *9*(7), e102670. https://doi.org/10.1371/journal.pone.0102670

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

Conner, M., & Norman, P. (2022). Understanding the intention-behavior gap: The role of intention strength. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.923464

Crook, M. N. (1941). Retest correlations in neuroticism. *Journal of General Psychology*, *24*, 173–182. https://doi.org/10.1080/00221309.1941.10544366

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

Darley, J. G. (1938). Changes in Measured Attitudes and Adjustments. *The Journal of Social Psychology*, *9*(2), 189–199. https://doi.org/10.1080/00224545.1938.9921688

Deary, I. J. (2014). The Stability of Intelligence From Childhood to Old Age. *Current Directions in Psychological Science*, *23*(4), 239–245. https://doi.org/10.1177/0963721414536905

Dohmen, T., Falk, A., Golsteyn, B. H. H., Huffman, D., & Sunde, U. (2017). Risk attitudes across the life course. *The Economic Journal*, *127*(605), F95–F116. https://doi.org/10.1111/ecoj.12322

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

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

Epper, T., Fehr, E., Hvidberg, K. B., Kreiner, C. T., Leth-Petersen, S., & Nytoft Rasmussen, G. (2022). Preferences predict who commits crime among young men. *Proceedings of the National Academy of Sciences*, *119*(6), e2112645119. https://doi.org/10.1073/pnas.2112645119

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Financial Services Authority. (2011). *Assessing suitability: Establishing the risk a customer is willing and able to take and making a suitable investment selection*. Financial Services Authority. https://www.fca.org.uk/publication/finalised-guidance/fsa-fg11-05.pdf

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

Frey, R., Richter, D., Schupp, J., Hertwig, R., & Mata, R. (2021). Identifying robust correlates of risk preference: A systematic approach using specification curve analysis. *Journal of Personality and Social Psychology*, *120*(2), 538–557. https://doi.org/10.1037/pspp0000287

Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, *1*(6), Article 6. https://doi.org/10.1038/s44159-022-00050-2

Gratton, C., Nelson, S. M., & Gordon, E. M. (2022). Brain-behavior correlations: Two paths toward reliability. *Neuron*, *110*(9), 1446–1449. https://doi.org/10.1016/j.neuron.2022.04.018

Green, L., Fry, A. F., & Myerson, J. (1994). Discounting of delayed rewards: A life-span comparison. *Psychological Science*, *5*(1), 33–36. https://doi.org/10.1111/j.1467-9280.1994.tb00610.x

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

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hertwig, R., Wulff, D. U., & Mata, R. (2019). Three gaps and what they may mean for risk preference. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1766), 20180140. https://doi.org/10.1098/rstb.2018.0140

Hopewell, S., Clarke, M. J., Stewart, L., & Tierney, J. (2005). Time to publication for results of clinical trials. *Cochrane Database of Systematic Reviews*, *2*. https://doi.org/10.1002/14651858.MR000011

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*(6), 543–549. https://doi.org/10.1016/j.jclinepi.2004.10.019

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

Kagaigai, A., & Grepperud, S. (2023). The role of risk preferences: Voluntary health insurance in rural Tanzania. *Health Economics Review*, *13*(1), 20. https://doi.org/10.1186/s13561-023-00432-z

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). Cumulative Meta-Analysis of Therapeutic Trials for Myocardial Infarction. *New England Journal of Medicine*, *327*(4), 248–254. https://doi.org/10.1056/NEJM199207233270406

Liu, S., Abdellaoui, A., Verweij, K. J. H., & van Wingen, G. A. (2023). Replicable brain–phenotype associations require large-scale neuroimaging data. *Nature Human Behaviour*, 1–13. https://doi.org/10.1038/s41562-023-01642-5

Malmendier, U., & Nagel, S. (2011). Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?*. *The Quarterly Journal of Economics*, *126*(1), 373–416. https://doi.org/10.1093/qje/qjq004

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), Article 7902. https://doi.org/10.1038/s41586-022-04492-9

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

Mata, R., Josef, A. K., & Hertwig, R. (2016). Propensity for Risk Taking Across the Life Span and Around the Globe. *Psychological Science*, *27*(2), 231–243. https://doi.org/10.1177/0956797615617811

Mata, R., Josef, A. K., Samanez-Larkin, G. R., & Hertwig, R. (2011). Age differences in risky choice: A meta-analysis. *Annals of the New York Academy of Sciences*, *1235*, 18–29. https://doi.org/10.1111/j.1749-6632.2011.06200.x

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/PR0.66.1.195-244

Mullen, B., Muellerleile, P., & Bryant, B. (2001). Cumulative Meta-Analysis: A Consideration of Indicators of Sufficiency and Stability. *Personality and Social Psychology Bulletin*, *27*(11), 1450–1462. https://doi.org/10.1177/01461672012711006

Mund, M., Freuding, M. M., Möbius, K., Horn, N., & Neyer, F. J. (2020). The stability and change of loneliness across the life span: A meta-analysis of longitudinal studies. *Personality and Social Psychology Review*, *24*(1), 24–52. https://doi.org/10.1177/1088868319850738

Murphy, B. A., & Lilienfeld, S. O. (2019). Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. *Psychological Assessment*, *31*(8), 1062–1072. https://doi.org/10.1037/pas0000732

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

Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*, *5*(11), 1535–1547. https://doi.org/10.1038/s41562-021-01117-5

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

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., … Schooler, J. (2020). High Replicability of Newly-Discovered Social-behavioral Findings is Achievable. *PsyArXiv*. https://doi.org/10.31234/osf.io/n2a9x

Santor, D. A., Gregus, M., & Welch, A. (2006). Eight Decades of Measurement in Depression. *Measurement: Interdisciplinary Research and Perspectives*, *4*(3), 135–155. https://doi.org/10.1207/s15366359mea0403_1

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

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

Seaman, K. L., Abiodun, S. J., Fenn, Z., Samanez-Larkin, G. R., & Mata, R. (2022). Temporal discounting across adulthood: A systematic review and meta-analysis. *Psychology and Aging*, *37*(1), 111. https://doi.org/10.1037/pag0000634

Sharma, L., Markon, K. E., & Clark, L. A. (2014). Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin*, *140*(2), 374–408. https://doi.org/10.1037/a0034418

Soutschek, A., & Tobler, P. N. (2018). Motivation for the greater good: Neural mechanisms of overcoming costs. *Current Opinion in Behavioral Sciences*, *22*, 96–105. https://doi.org/10.1016/j.cobeha.2018.01.025

Sparrow, E. P., Swirsky, L. T., Kudus, F., & Spaniol, J. (2021). Aging and altruism: A meta-analysis. *Psychology and Aging*, *36*(1), 49–56. https://doi.org/10/gj2gsj

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, *117*(30), 17680–17687. https://doi.org/10.1073/pnas.1920484117

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

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

United Nations, Department of Economic and Social Affairs, Population Division. (2019). *World Population Ageing 2019: Highlights*. (ST/ESA/SER.A/430).

Westbrook, A., Kester, D., & Braver, T. (2013). What Is the Subjective Cost of Cognitive Effort? Load, Trait, and Aging Effects Revealed by Economic Preference. *PloS One*, *8*, e68210. https://doi.org/10.1371/journal.pone.0068210

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

Yang, M., Roope, L. S. J., Buchanan, J., Attema, A. E., Clarke, P. M., Walker, A. S., & Wordsworth, S. (2022). Eliciting risk preferences that predict risky health behavior: A comparison of two approaches. *Health Economics*, *31*(5), 836–858. https://doi.org/10.1002/hec.4486

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why Current Publication Practices May Distort Science. *PLOS Medicine*, *5*(10), e201. https://doi.org/10.1371/journal.pmed.0050201

Appendices

Appendix A: Manuscript 1

Bagaïni, A., Liu, Y., Kapoor, M., Son, G., Bürkner, P.C., Tisdall, L., & Mata, R. (2023). Comparing the Temporal Stability and Convergent Validity of Risk Preference Measures: A Meta-Analytic Approach. *Manuscript Submitted for Publication.*

Appendix B: Manuscript 2

Bagaïni, A., Liu, Y., Bajrami, A., Son, G., Tisdall, L., & Mata, R. (2023). Aging and Economic Preferences: Cumulative meta-analyses of age differences in risk, time, social, and effort preferences. *The Journals of Gerontology: Series B, 78*(8), 1122-1135. doi: 10.1093/geronb/gbad034

Appendix C: Manuscript 3

Liu, Y., Bagaïni, A., Son, G., Kapoor, M., & Mata, R. (2023). Life-course trajectories of risk-taking propensity: A coordinated analysis of longitudinal studies. *The Journals of Gerontology: Series B, 78*(3), 445-455. doi:10.1093/geronb/gbac175

Appendix A

Manuscript 1

Bagaïni, A., Liu, Y., Kapoor, M., Son, G., Bürkner, P.C., Tisdall, L., & Mata, R. (2023). Comparing the Temporal Stability and Convergent Validity of Risk Preference Measures: A Meta-Analytic Approach. *Manuscript Submitted for Publication.*

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

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 S1 also 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 traits. 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 Figure5, 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 (propensity, frequency) or psychological constructs 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. 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.

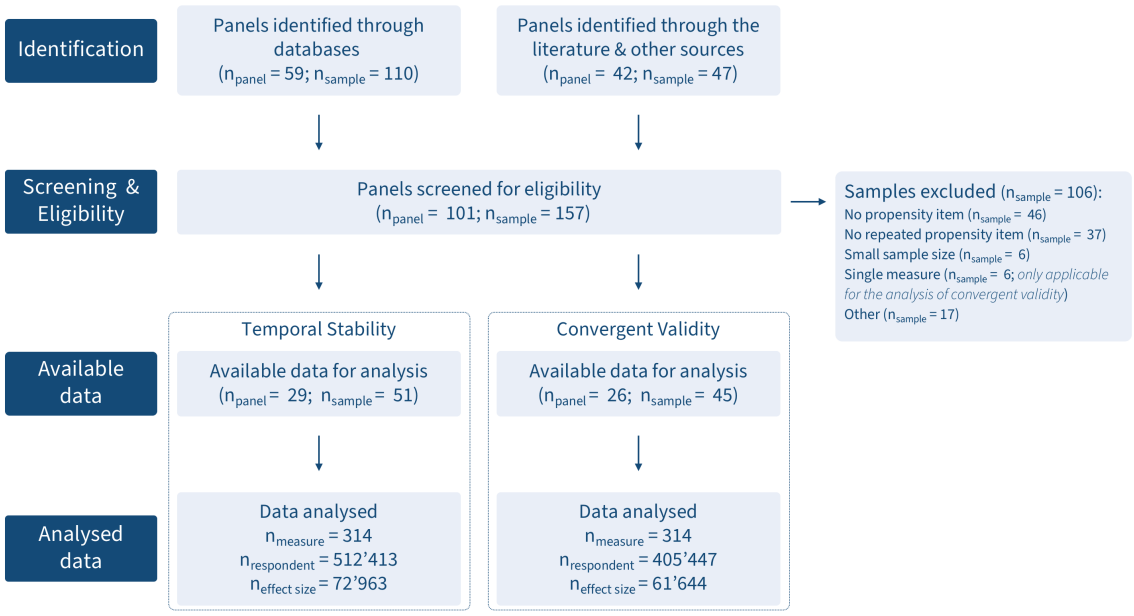


Figure 2

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.

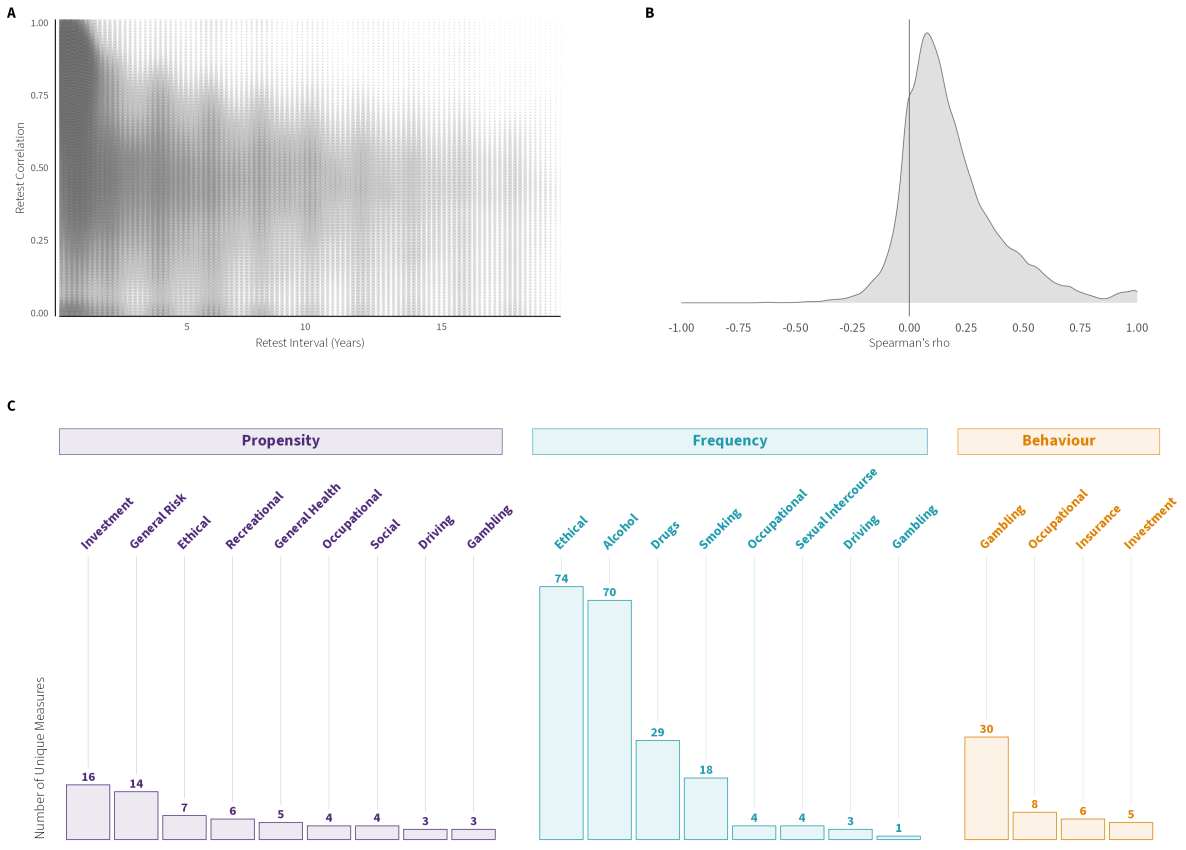


Figure 3

Variance decomposition of temporal stability. A) Relative contribution of measure, respondent, and panel 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 mean estimate, shaded uncertainty bands represent 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)

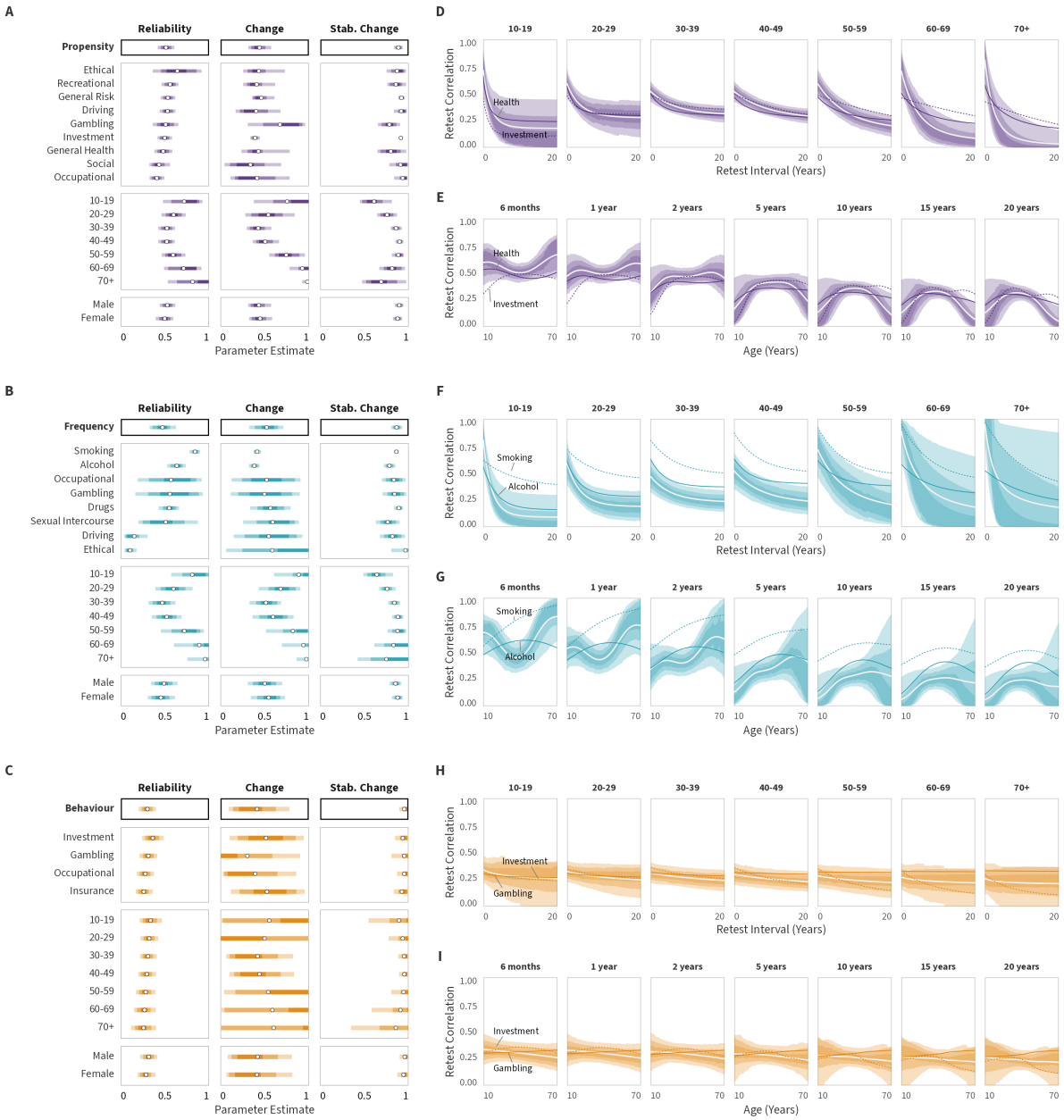


Figure 5

Variance decomposition of convergence between measures. Relative contribution of measure, respondent, and panel 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.

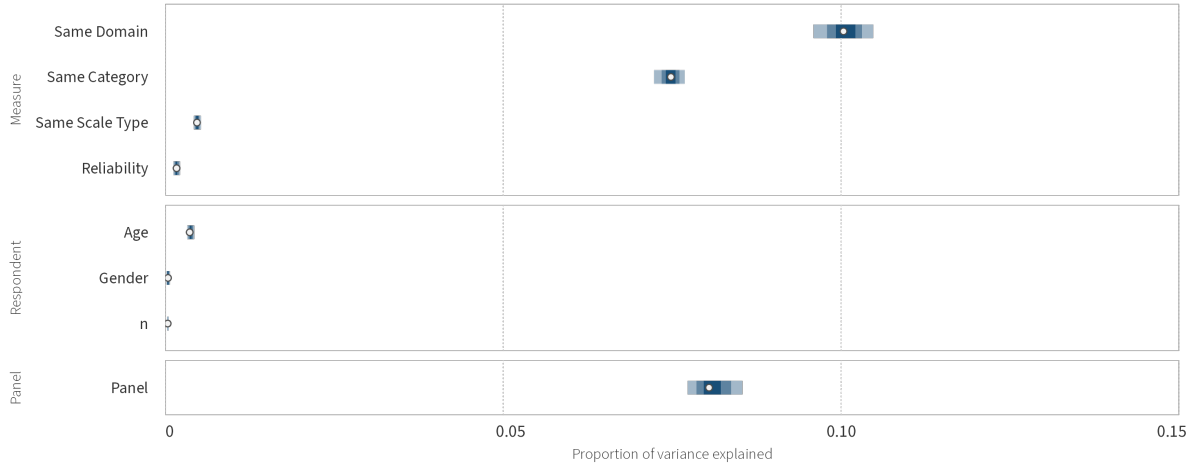


Figure 6

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

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 type 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 sample 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).

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

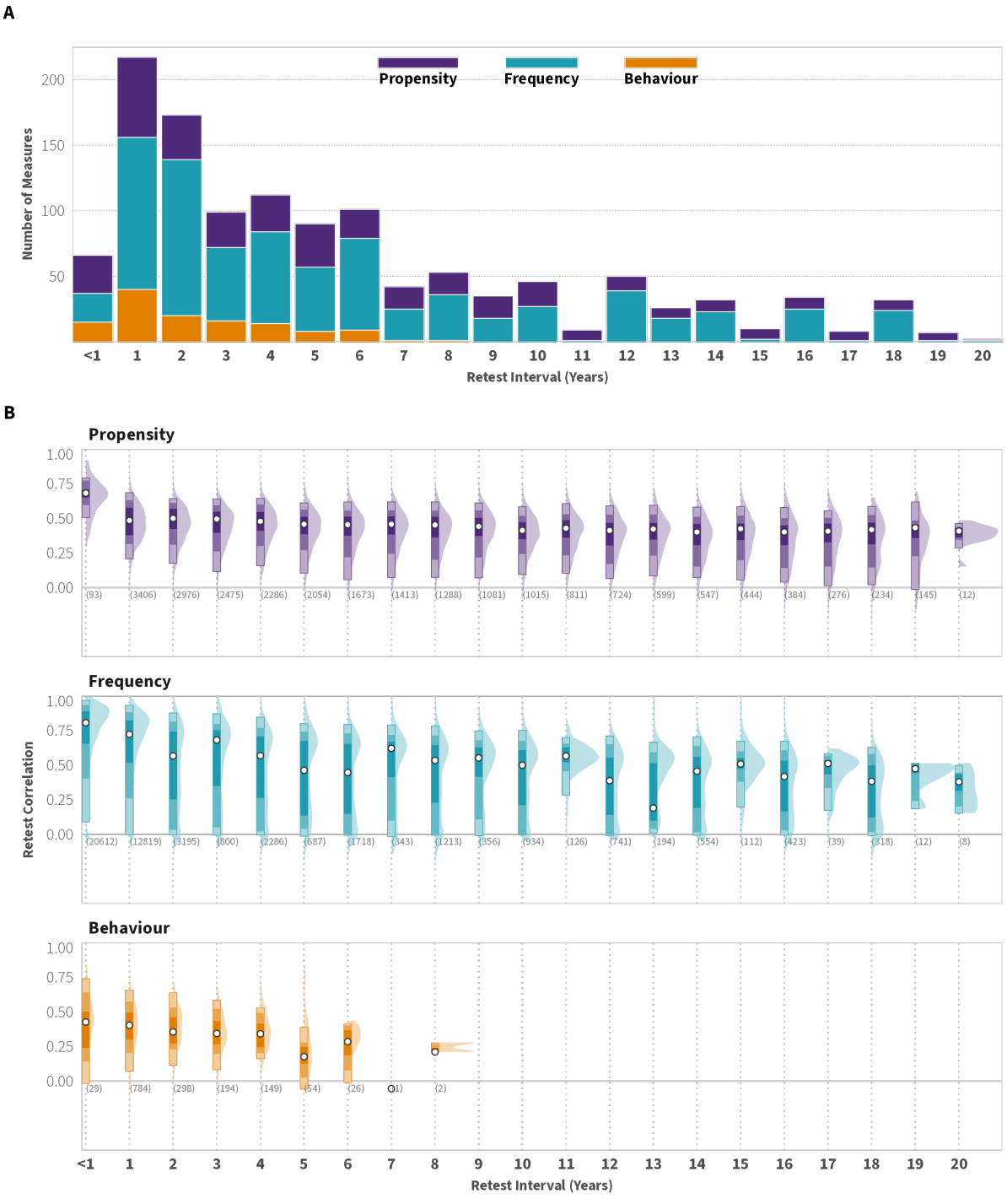
In both meta-regressions we specified predictors for the residual standard deviations, and allowed it to vary across the different levels of the categorical variables. Based on recommendations, in all models, we used weakly informative priors (Williams et al., 2018). Lastly, we back-transformed the results to Pearson's r 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 raw retest correlations as a function of retest interval for the different measure categories (propensity, frequency, behaviour).



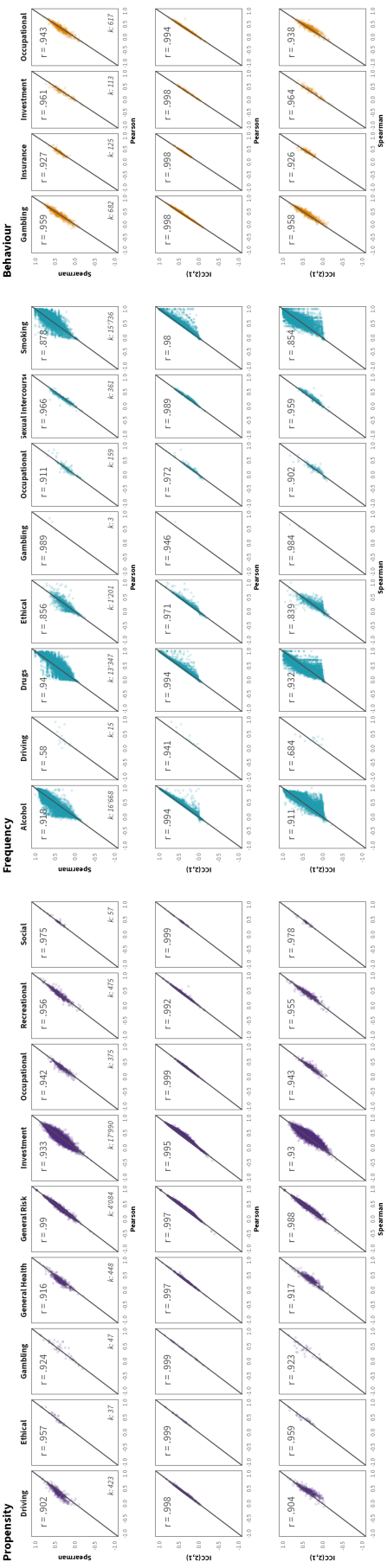


Figure S2

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

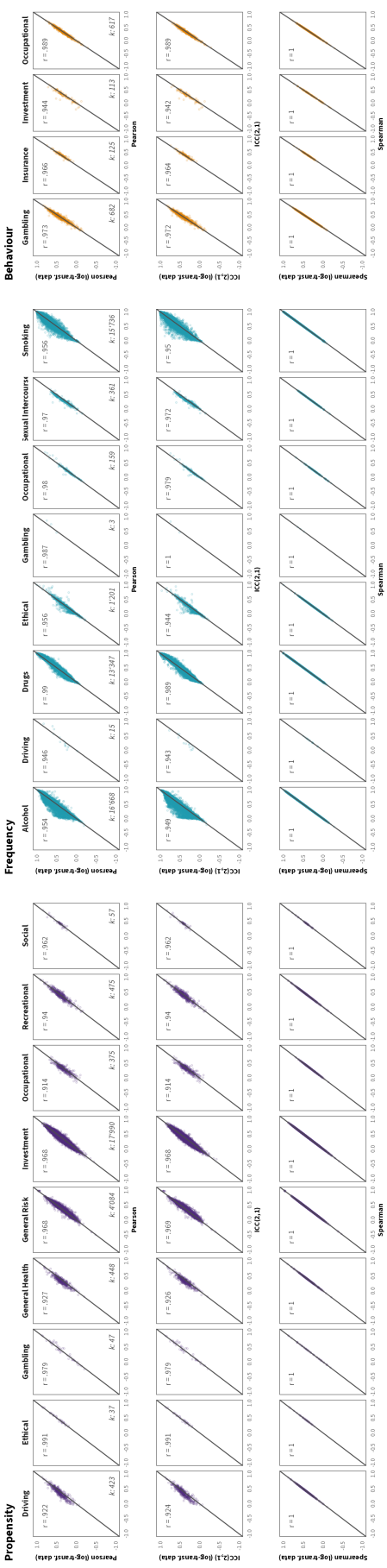


Figure S3

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

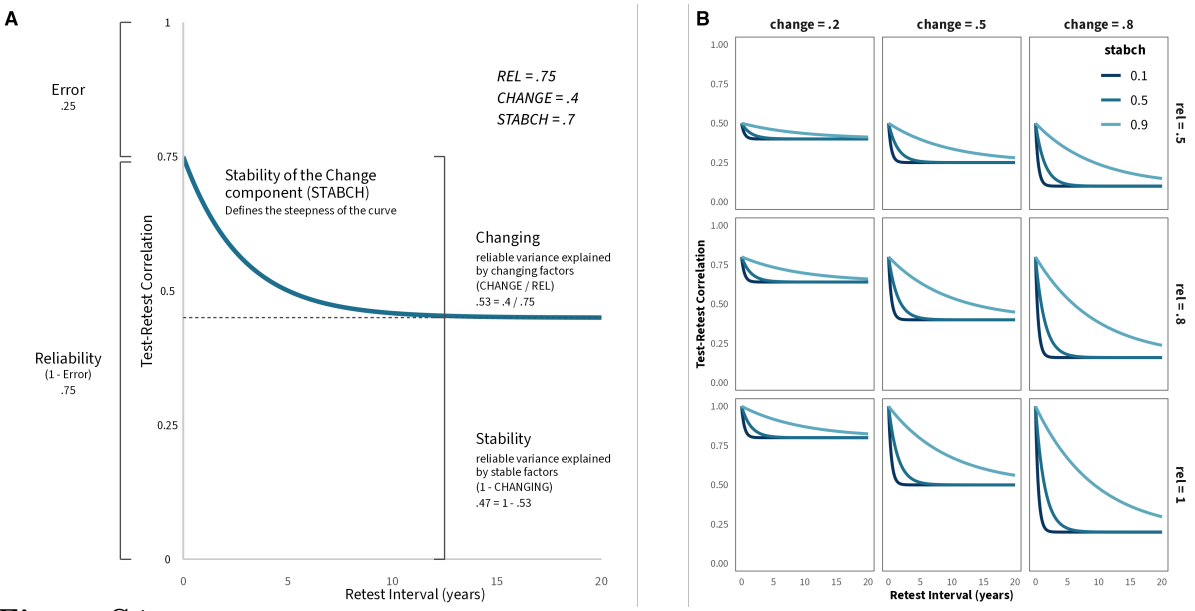


Figure S4

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 & Schimmack, 2016). B) Examples of different parameterisations of MASC.

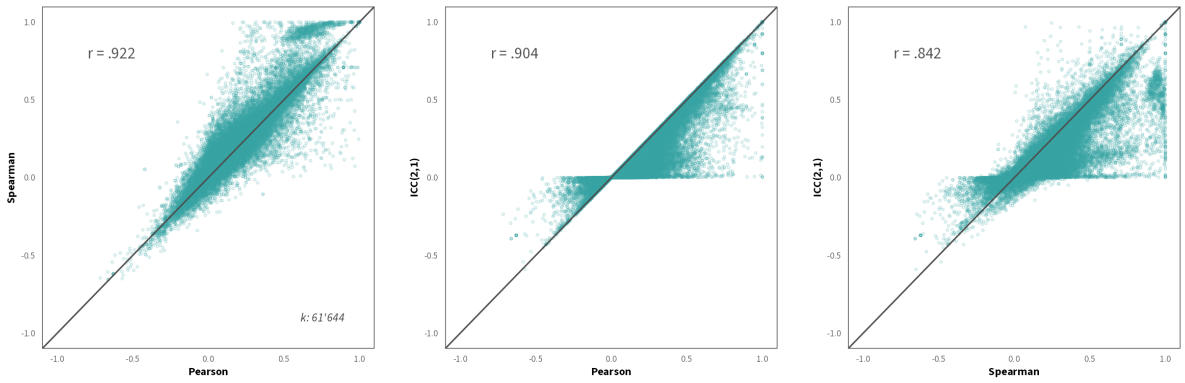


Figure S5

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

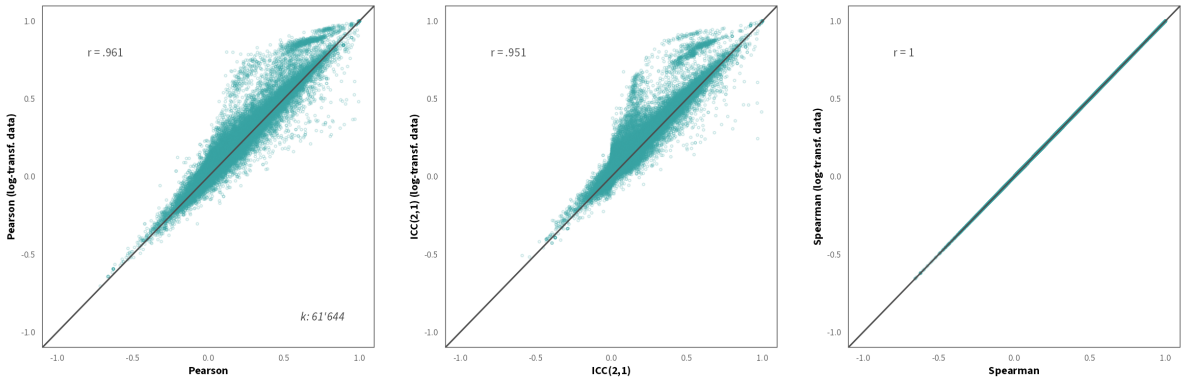


Figure S6

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).

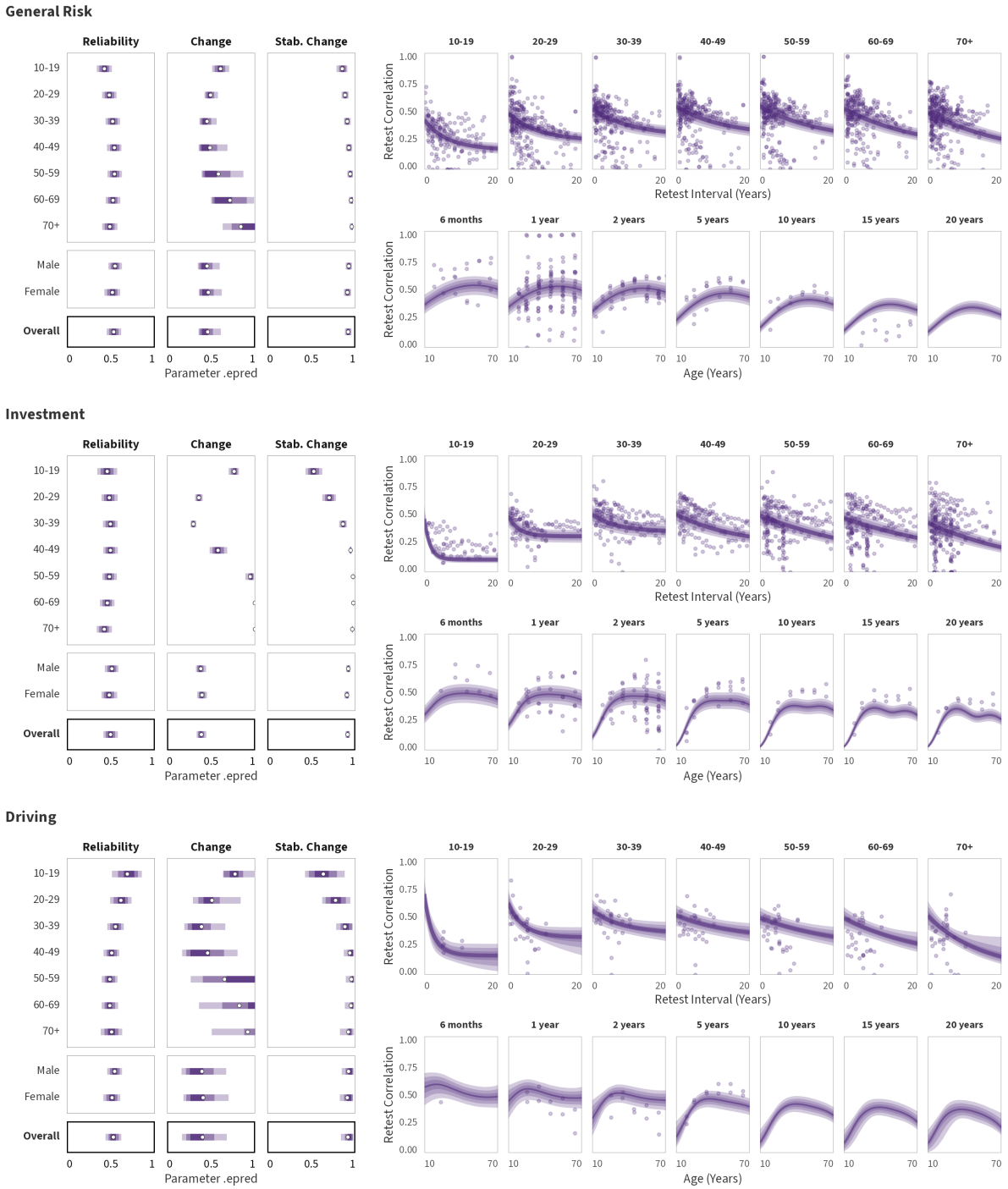


Figure S8

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).

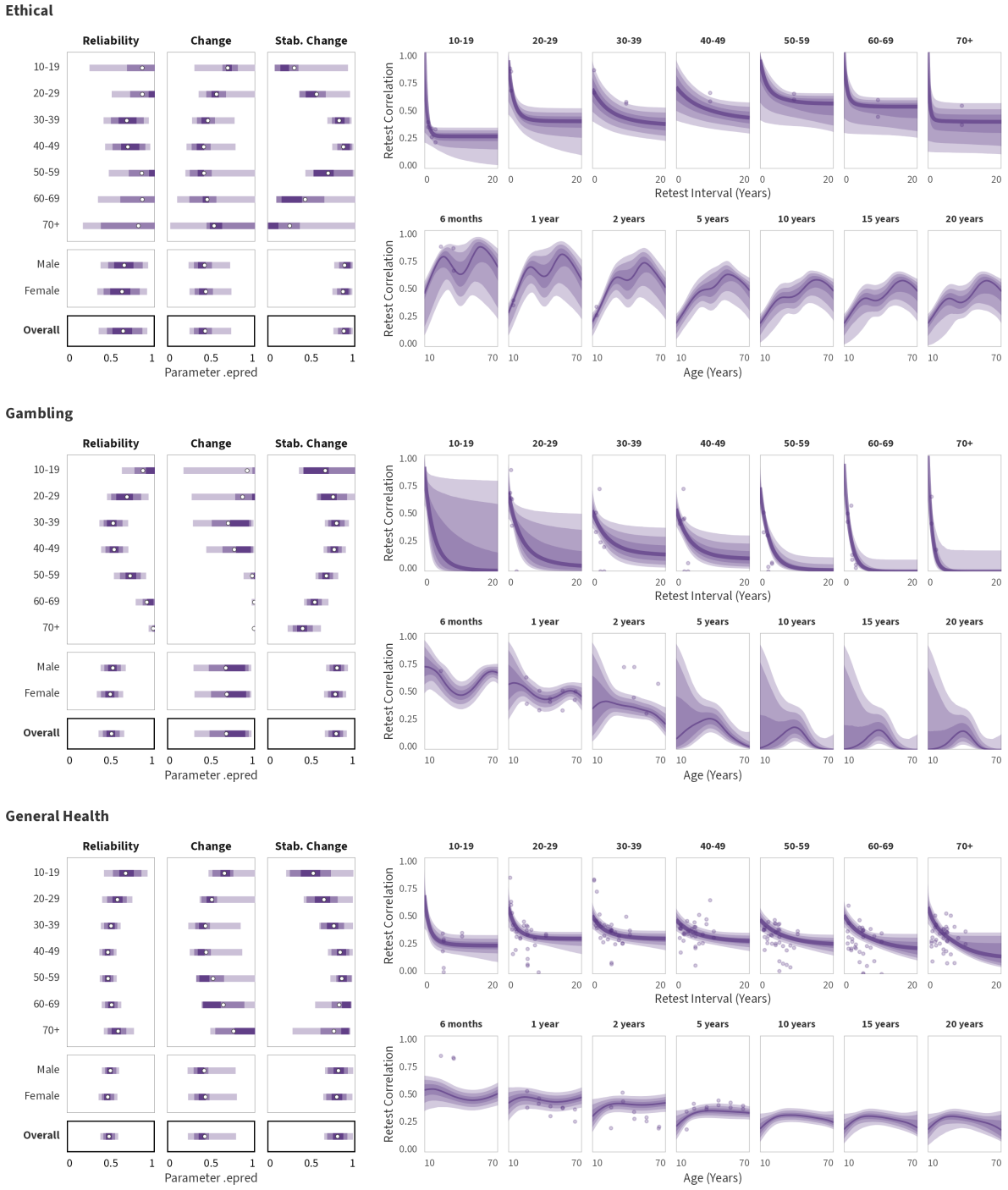


Figure S9

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).

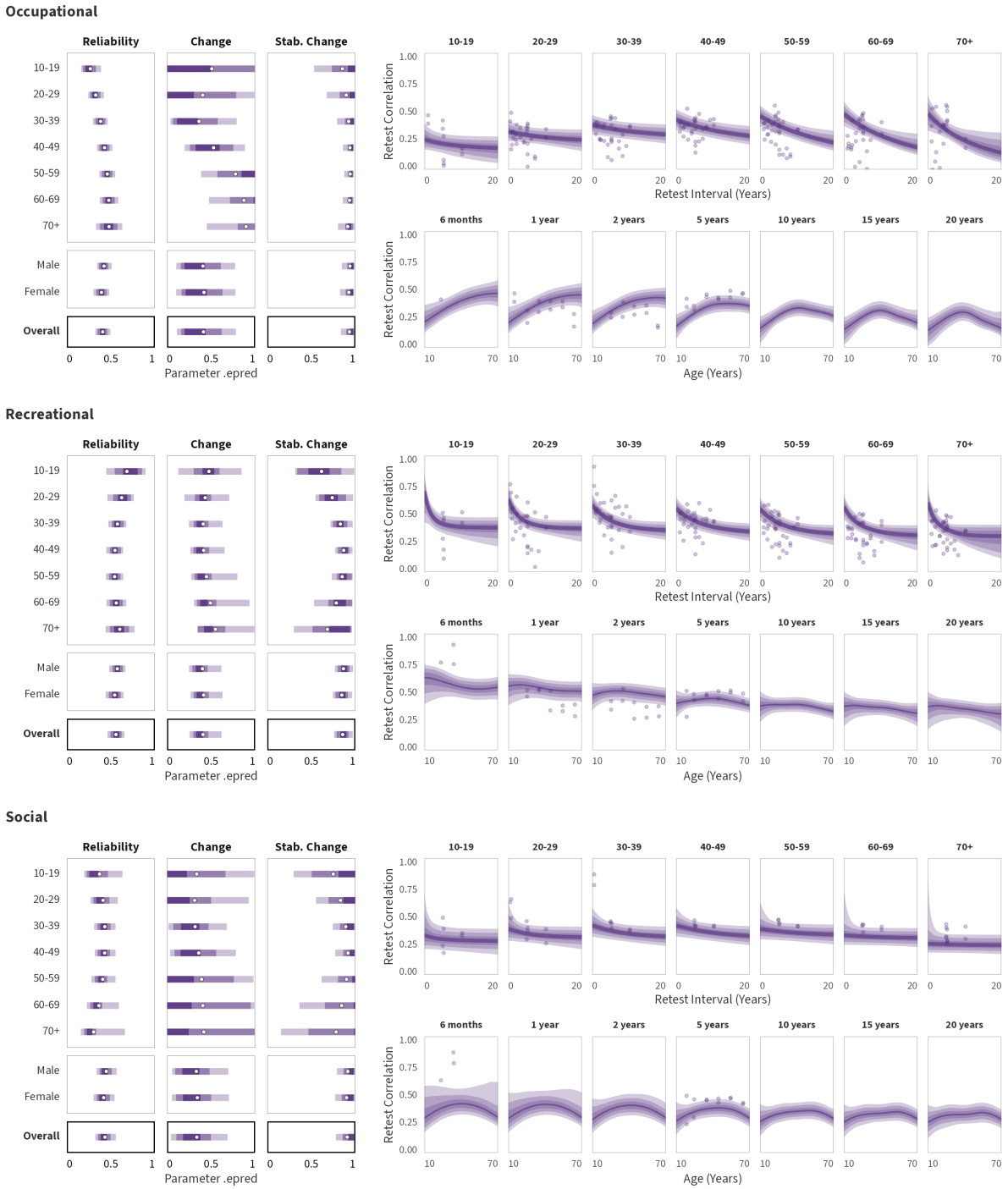


Figure S10

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).

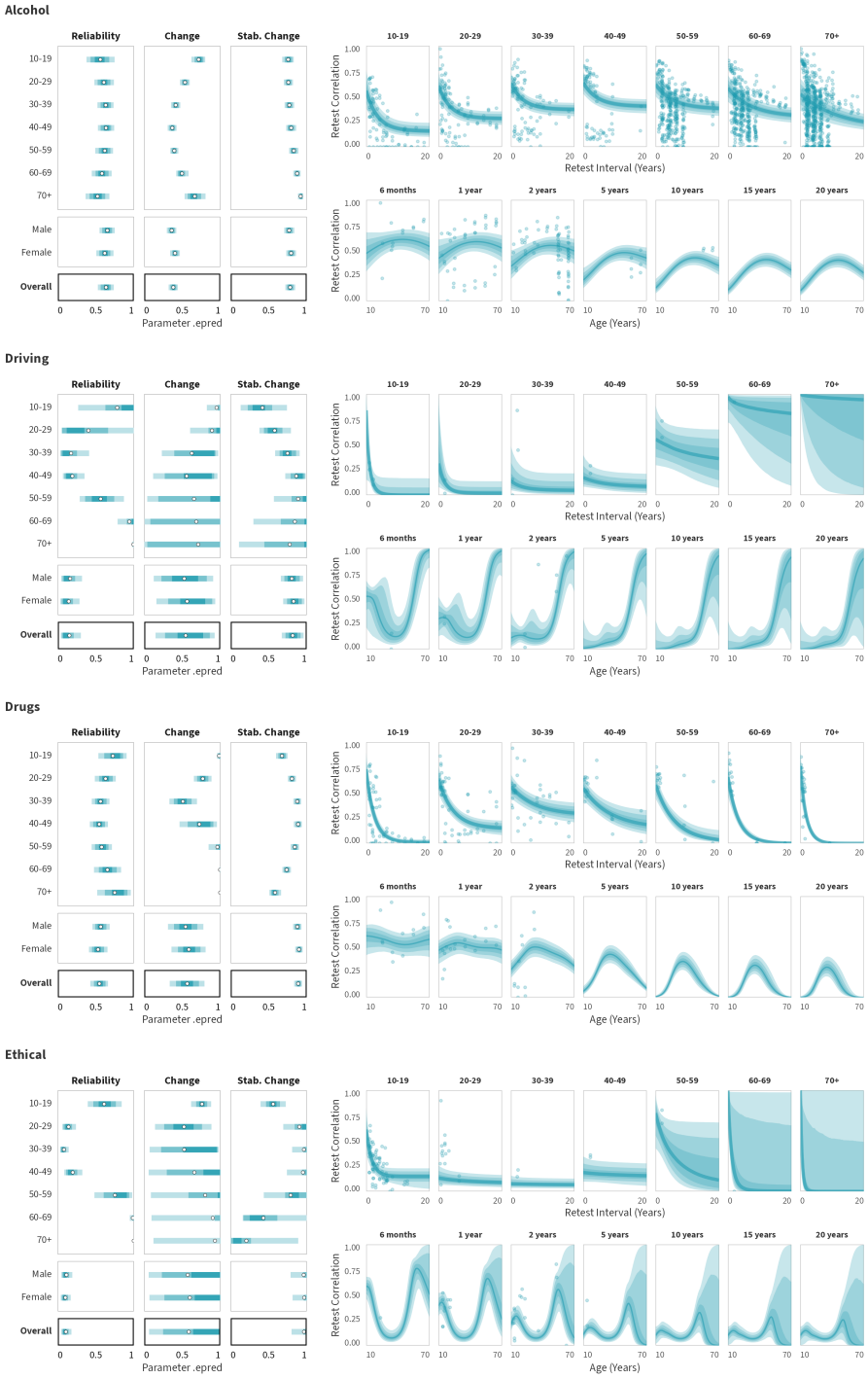


Figure S11

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).

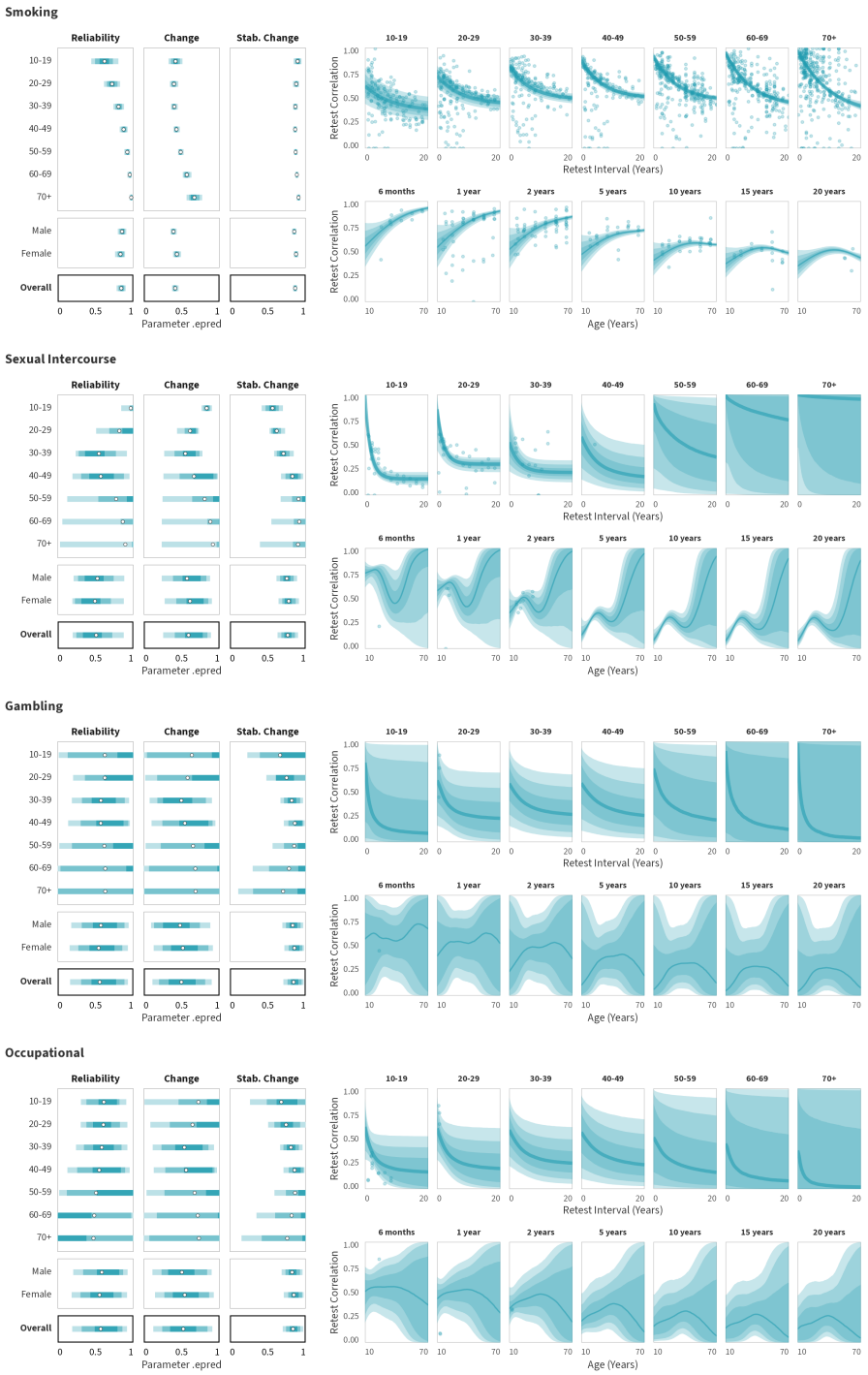


Figure S12

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).

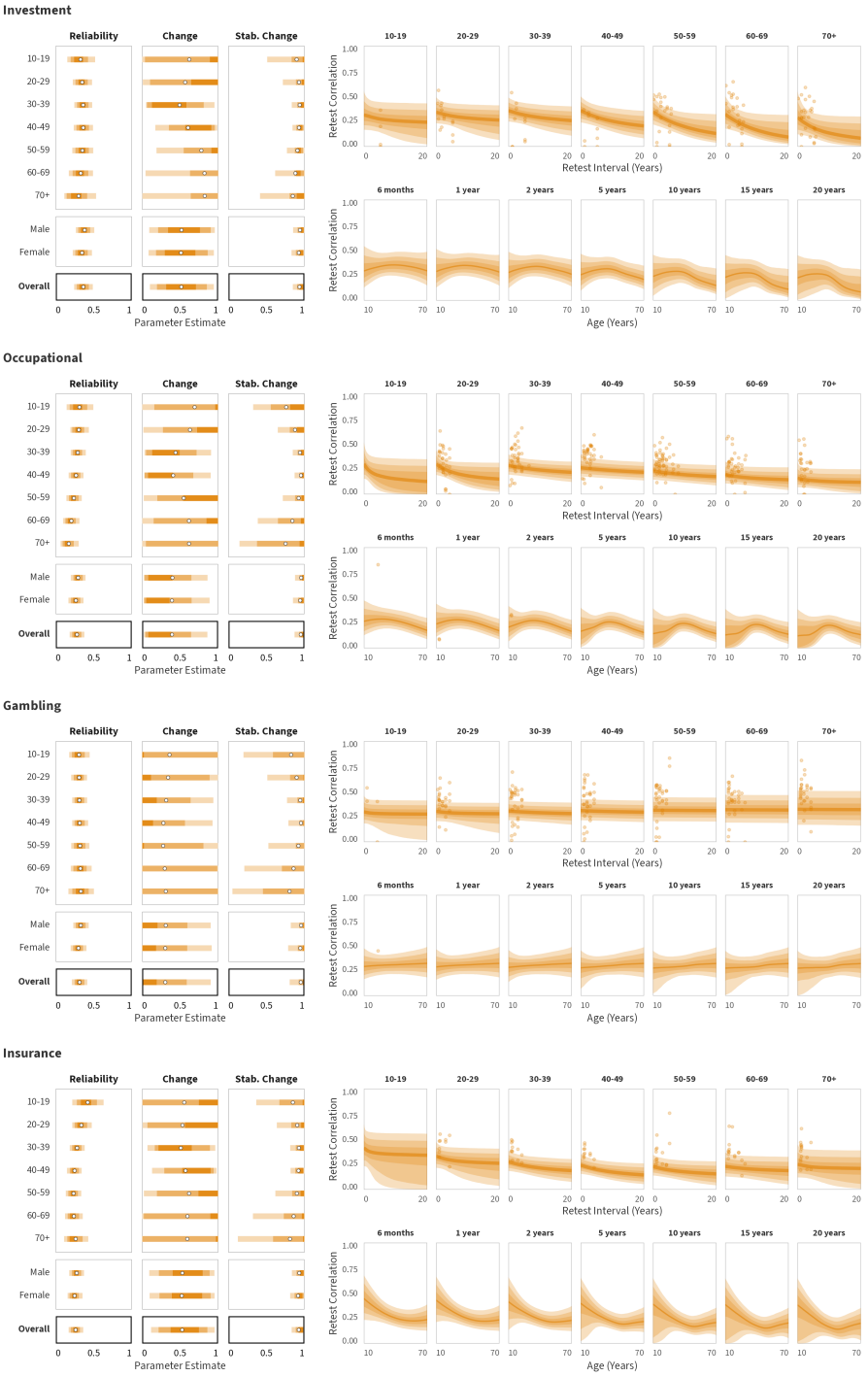


Figure S13

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).

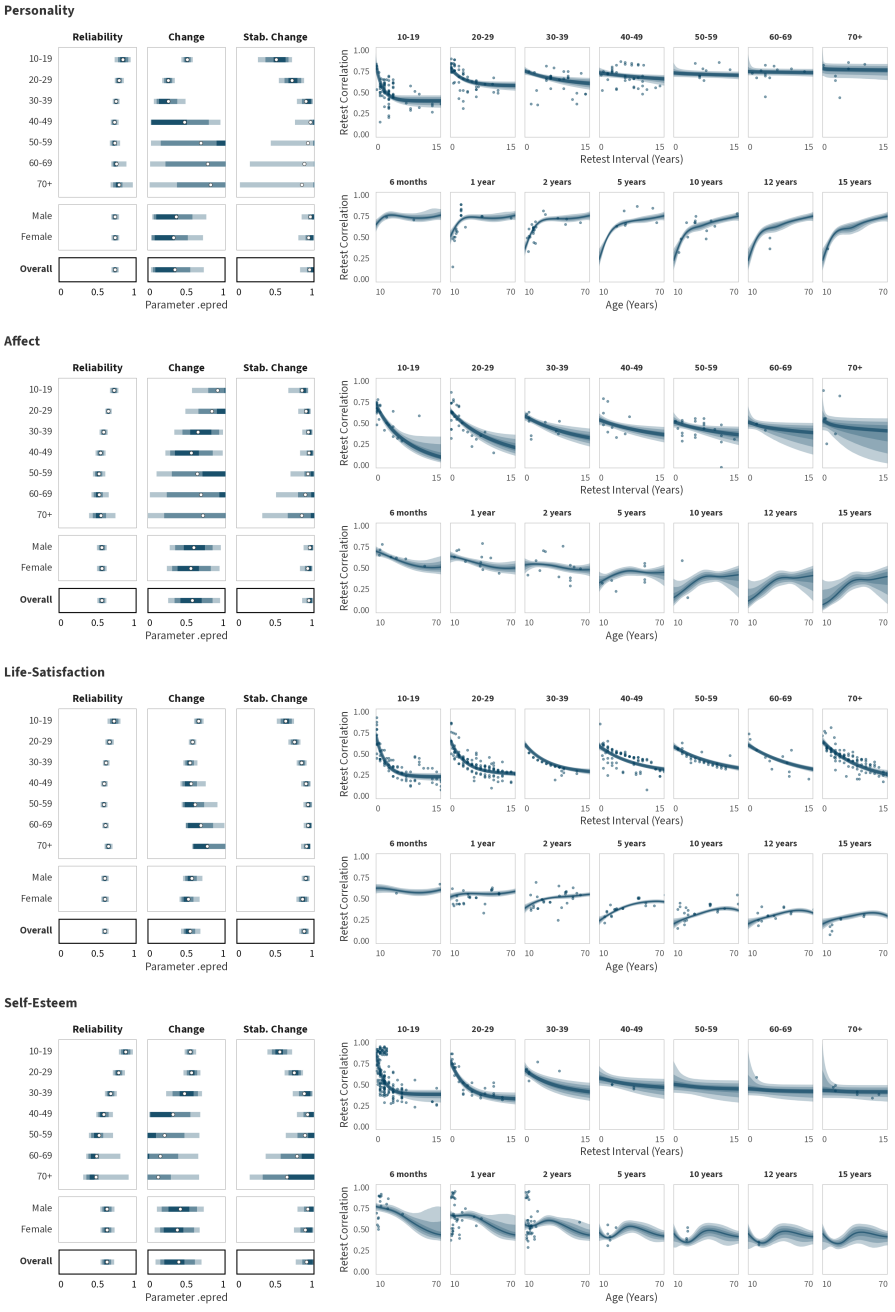


Figure S14

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).

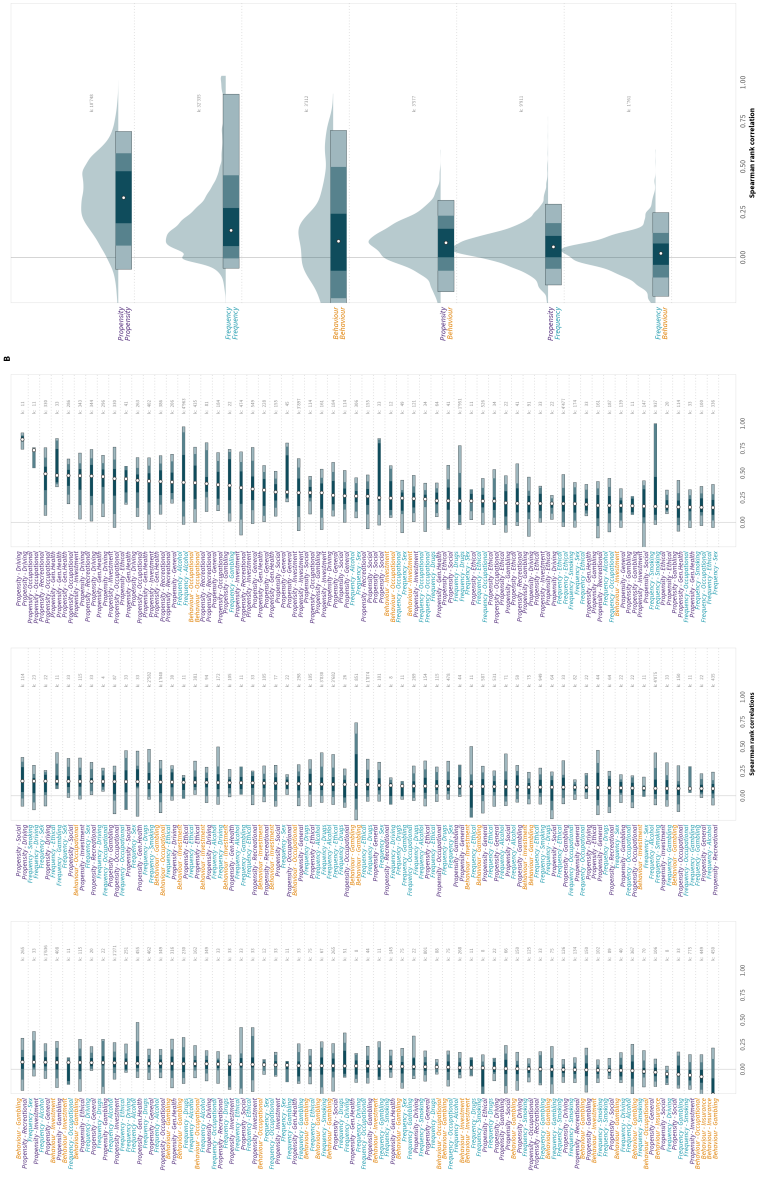


Table S1

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

Appendix B

Manuscript 2

Bagaïni, A., Liu, Y., Bajrami, A., Son, G., Tisdall, L., & Mata, R. (2023). Aging and Economic Preferences: Cumulative meta-analyses of age differences in risk, time, social, and effort preferences. *The Journals of Gerontology: Series B, 78*(8), 1122-1135. doi: 10.1093/geronb/gbad034

Aging and Economic Preferences:

Cumulative meta-analyses of age differences in risk, time, social, and effort preferences

Alexandra Bagaïni, M.Sc.[1], Yunrui Liu, M.Sc.[1], Arzie Bajrami, B.Sc.[1], Gayoung Son, B.Sc.[1], Loreen Tisdall, Ph.D.[1], and Rui Mata, Ph.D.[1]

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

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

Objectives: Several theories predict changes in individuals' economic preferences across the life span. To test these theories and provide an historical overview of this literature, we conducted meta-analyses on age differences in risk, time, social, and effort preferences as assessed by behavioral measures.

Methods: We conducted separate meta-analyses and cumulative meta-analyses on the association between age and risk, time, social, and effort preferences. We also conducted analyses of historical trends in sample sizes and citations patterns for each economic preference.

Results: The meta-analyses identified overall no significant effects of age for risk (r = -0.02, 95%CI[-0.06, 0.02], n = 39,832), and effort preferences (r = 0.24, 95%CI[-0.05, 0.52], n = 571), but significant effects of age for time (r = -0.04, 95%CI[-0.07, -0.01], n = 115,496) and social preferences (r = 0.11, 95%CI[0.01, 0.21], n = 2,997), suggesting increased patience and altruism with age, respectively. Equivalence tests, that compare these effects to practically important ones (i.e., r = |.1|), however, suggest that all effects are of trivial significance. The analyses of temporal trends suggest that the magnitude of effects and sample sizes have not changed significantly over time, nor do they dramatically affect the extent that articles are cited.

Discussion: Overall, our results contrast with theories of aging that propose general age effects for risk, and effort preferences, yet provide some but tenuous support for those suggesting age-related changes in time and social preferences. We discuss implications for theory development as well as future empirical work on economic preferences.

keywords: cumulative, meta-analysis, age differences, economic preferences

Aging and Economic Preferences:

Cumulative meta-analyses of age differences in risk, time, social, and effort preferences

Aging is thought to be associated with changes in decision-making that can carry long-term consequences for oneself as well as others, including choices about financial investment, savings, donations, or effort expenditure. Economic preferences reflect how individuals tend to make associated trade-offs about risk, time, social, or effort dimensions when making such choices and there has been considerable interest in understanding how and to what extent such preferences change with age (e.g., Best & Charness, 2015; Seaman et al., 2022; Sparrow et al., 2021; Westbrook et al., 2013). The empirical results concerning economic preferences have, however, been mixed and there have been recent calls to examine the research practices associated with aging research and harmonizing both theories and methods to advance the study of age differences in economic preferences (e.g., Frey et al., 2021). For example, some researchers have voiced concern about a potential tendency to exclusively report significant age differences in the aging literature (e.g., Isaacowitz, 2018, 2020) or how certain stylized facts about the link between aging and economic preferences may reflect the work of a few seminal studies that are based on relatively small sample sizes and are not representative of the literature as a whole (e.g., Seaman et al., 2022).

In this work, we aim to contribute to integrating both theory and empirical knowledge about age differences in economic preferences by providing a comprehensive research synthesis of this literature to assess how the different existing theories in this domain match with the empirical evidence accumulated over time. Taking stock of the amount and time course of how evidence accumulates over time can provide insights into the history of the field, the impact of evolving research practices (e.g., study designs, sample sizes, statistical approaches), and the stability of the knowledge acquired (Koricheva et al., 2013; Kulinskaya & Mah, 2022). We thus aim to provide an overall assessment of how different theories of

aging are supported or rejected by current empirical evidence and provide input for both theory development and future empirical work in the domain of economic preferences.

Economic Preferences: Risk, Time, Social and Effort

In this study, we focus on age-related differences in four domains of economic preference: risk, time, social, and effort-related preferences. Table 1 provides a summary of these constructs along with examples of tasks commonly used in the psychological literature to assess them. Risk preference can be defined as the propensity of an individual to prefer options offering varying (monetary) rewards over certain ones. Popular tasks involve choosing between gambles of varying levels of (learned or described) rewards or probabilities (e.g., Holt and Laury, 2002). Time preference is defined by how much an individual discounts the value of future rewards over sooner ones. Most often a discounting rate is estimated based on the choices an individual makes between immediate rewards and larger delayed rewards in a temporal discounting task (Frederick et al., 2002). Social preference reflects an individual's inclination to forgo resources for oneself for the sake of another individual. The dictator game (Forsythe et al., 1994) is a commonly used task where a player chooses to donate a certain amount of real or hypothetical money to an undisclosed participant. Lastly, effort preferences are typically conceptualized as effort discounting and calculated by how much the subjective value of a reward decreases as a function of the cognitive or physical effort needed to acquire it (e.g., Ostaszewski et al., 2013).

Theoretical Accounts Predicting Age Differences in Economic Preferences

As outlined earlier, a number of theoretical approaches have made predictions about the life-span development of the economic. In what follows, we discuss a number of such theories with a particular focus on those that have been used to make predictions across different types of economic preferences.

Socio-emotional selectivity theory (cf., Carstensen, 2006) is a prominent motivational theory that has been used to derive prediction across a number of economic preferences, including risk, time, and social preferences. It postulates that with age, individual's future time horizon shrinks, which results in a shift in goal orientation, from future- to present-oriented as well as from the self to others. There has been some discussion about the empirical status of socio-emotional selectivity theory and how it can be distinguished from other motivational theories (cf. Depping & Freund, 2011) but there seems to be some consensus that the theory predicts a decrease in risk taking, increased temporal discounting, as well as increased altruism with age (cf. Frey et al., 2021; Seaman et al., 2022; Sparrow et al., 2021).

Other theories have proposed that age differences in economic decisions can be the result of relatively general neurological changes. For example, the dopaminergic neuromodulation hypothesis posits that a decline in dopaminergic functioning reduces older adult's responses towards rewards. Therefore, older adults in comparison to younger adults are less motivated to obtain rewards, leading to a reduction in the propensity to take risks or exert effort to obtain a larger reward, as well as a decreased need to obtain an immediate reward (cf. Frey et al., 2021; Seaman et al., 2022; Westbrook et al., 2013).

Other theories consider the interaction between age-related cognitive decline and task characteristics. Specifically, the confound hypothesis suggests that there may be differences between types of tasks as a function of their cognitive demands and different aspects of cognitive functioning (e.g., fluid vs. crystallized aspects) that can moderate age effects (cf., Mata et al., 2011). This is particularly applicable to risk and time preference tasks, in which researchers have shown that estimates for risk preference and temporal discounting can appear to increase or decrease as a function of task demands or analytic confounds (cf. Frey, Mata, & Hertwig, 2015; Olschewski et al., 2018).

Other theories have focused on other non-psychological causes that covary with age, such as socio-economic factors, including one's social network and financial wealth, that can shape individuals' economic preferences. For example, the accumulation of social and economic capital implies reduced striving for such resources across the life span, leading to changes in financial risk-taking and social behavior (cf., Frey et al., 2021; Mayr & Freund, 2020).

Aside from theories that make predictions across several economic preferences, a number of preference-specific theories have also been advanced. For example, for risk preference, evolutionary signalling theory presents risk taking as an indication of fitness that is most relevant to younger adults that need to signal fitness for reproductive reasons (cf. Frey et al., 2021). For time preference, some have suggested that the perception of time changes, whereby with age, the impression that time goes by more quickly becomes more common, which can reduce the perception of amount of time to wait to obtain a larger reward, and in turn increases one's willingness to wait (cf., Seaman et al., 2022). Concerning prosocial behaviour, the intuitive-prosociality hypothesis describes altruism as an intuitive response that tends to increase with age (cf., Mayr & Freund, 2020). Lastly, regarding effort, selective engagement theory postulates that age-related increases in the perception of costs related to a task decreases the willingness to expend effort (cf. Hess et al., 2021).

All in all, this short survey highlights the rather heterogenous character of theories, spanning motivational, cognitive, and ecological factors, and the plethora of mechanisms proposed in the past literature. Table 2 summarizes these and other theoretical accounts and lists relevant references to provide an overview of predictions about how age is associated with each of the four types of economic preferences. In this paper we examine the match between these predictions and the empirical evidence across types of preferences in a systematic fashion. Such integrative efforts are important as they provide an assessment of

the scope of theories and help us gain a better sense of their strengths and limitations. Additionally, examining a theory across multiple domains can help identify inconsistencies or gaps, and provide insight into how the theory can be refined or expanded.

Past Empirical Evidence

Given the diversity of theoretical approaches in place, it may not be surprising that existing reviews and meta-analyses on the effect of age on risk, time, or social preferences report findings that are not fully consistent with all the proposed theories (Best & Charness, 2015; Mata et al., 2011; Seaman et al., 2022; Sparrow et al., 2021). In the context of age-related differences in risk preference, the most recent meta-analysis of behavioral measures found no overall effect of age on risk preference but reported that age differences depend on their context (health vs. monetary) and domain, specifically, gains versus losses (Best & Charness, 2015). An earlier meta-analysis also found no overall effect of age on risk preference but did report suggestive evidence that age differences may be evident for tasks that involve learning from experience (Mata et al., 2011). For time preference, a recent meta-analysis reported no significant main effect of age (Seaman et al., 2022). In line with theory, a recent meta-analysis that synthesized evidence on age-related differences in social preference involving a mix of measure types (behavioral tasks, self-reports) reported a medium-sized effect of age, with older adults showing greater altruistic tendencies than younger adults (Sparrow et al., 2021). Finally, thus far, no meta-analysis has been conducted on age differences in effort discounting, but primary studies show conflicting results regarding age differences (e.g., Hess et al., 2021; Seaman et al., 2016).

Despite the past empirical work including research synthesis in this area, it is still difficult to adequately compare the empirical results to theories for several reasons. First, each meta-analysis captured the state of the literature at a specific point in time and thus may have captured different amounts and types of evidence that bear on the theories in question.

Second, the meta-analyses did not share the same eligibility criteria, such as sample characteristics, study designs, or types of measures (behavior vs. self-report). Third, more broadly, past syntheses have not assessed how evidence on age effects accumulated over time and to what extent changing research practices such as the introduction of specific paradigms or study characteristics (sample size, study context) have influenced the estimates of age differences in economic preferences or their impact. We believe, however, that putting our estimates of age effects in an historical context could be important to either assuage or strengthen concerns about the status of the aging literature (e.g., Isaacowitz, 2020).

Overview of the Current Study

In this study, we aim to address limitations of past work by offering an updated overview of age effects in risk, time, social, and effort preferences. We focus specifically on studies that have investigated age differences in economic preferences as measured through behavioral tasks involving financial decisions. The main rationale for focusing on behavioral measures in the financial domain is to maximize comparability across types of economic preferences. This is important because recent work suggests that different measures types (behavioral measures vs. self-reports) do not always produce similar results concerning age effects in economic preferences (e.g., Frey et al., 2021). Consequently, in our work, we update and harmonize previous meta-analyses by focusing specifically on behavioral tasks in the financial domain. Relatedly, this also allows us to explore the role of a large range of theoretically and empirically motivated moderators across all preferences (see Table S1 for an overview). Further, we extend past syntheses by conducting cumulative meta-analyses to gain insight into how estimates of age effects changed over time as evidence accumulated in the literature. Cumulative meta-analysis is the process of updating meta-analytic results by incorporating new evidence (Lau et al., 1992) and this approach can help detect historical trends, evaluate evidence sufficiency, and possibly identify selective reporting, such as time-

lag bias or the Proteus phenomenon (i.e., the tendency for early replications of a scientific work to contradict the original findings; Ioannidis & Trikalinos, 2005; Koricheva et al., 2013; Young et al., 2008), which has been implied in past aging work (Seaman et al., 2022). Lastly, some areas of psychology have seen noticeable changes over time that are linked to new research practices (e.g., conducting online studies) that allow for convenient sampling of larger samples and can have consequences for the quantity and quality of data (Sassenberg & Ditrich, 2019). Consequently, we explore the link between time of publication and sample sizes as a way to assess whether research practices have changed over time in the context of economic preferences, as well as assess studies' impact by analysing their historical citation patterns. Overall, we hope to determine the robustness and stability of estimates of age effects in economic preferences so as to be able to draw robust conclusions about the match between the observed empirical patterns and extant theoretical predictions.

Method

Our research synthesis approach involved two steps. First, we conducted a scoping review of the aging literature to identify existing meta-analyses that have estimated age differences in economic preferences (see the Supplementary Appendix for details on our search strategy and results). Our main goal was to make sure we included all eligible primary studies from these existing reviews. Second, we performed a search for additional primary studies following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021) with the goal of complementing the coverage of past research syntheses. Below we describe the steps involved in the search, screening and data extraction for primary studies on age differences in risk, time, social, and effort preferences.

Literature Search

For time preference, we complemented the list of primary studies from the Seaman et al. (2022) meta-analysis with more recent studies whereas for risk, social, and effort preferences, we conducted whole new searches (for papers published until November 1st, 2022; Table S2). We did not complement previous meta-analyses (Best & Charness, 2015; Mata et al., 2011; Sparrow et al., 2021) due to significant differences with the eligibility criteria, analysis, and coding used by Seaman et al. (2022).

The searches returned 2052, 315, 460 and 510 candidate studies to screen for risk, time, social and effort preferences, respectively.

Screening

To screen the articles resulting from the search, we devised a set of criteria that we harmonized and applied across all four preferences. We used the same criteria as Seaman et al. (2022), with the exception that we excluded unpublished studies (that would be difficult to place in an historical analysis) and studies that collected data while participants underwent brain imaging, brain stimulation, or pharmacological studies (that would decrease comparability). An overview of the general and preference-specific criteria is available in Table S3.

From the search results, we first screened studies based on the title and abstract, and removed 1817, 232, 414, and 441 studies for risk, time, social, and effort preference, respectively. Individual study members then reviewed the remaining full-text articles. We observed that certain articles that we included for the analysis, investigated multiple economic preferences. Therefore, we complemented the list of included articles across preferences by adding articles that had been included in the meta-analysis of one economic preference and met the criteria of another but that had not been identified in the search of this one. In the end, a total of 57, 50, 13, and 6 published articles were included in the analysis of

age differences in risk, time, social, and effort preference, respectively. The process is illustrated in separate PRISMA flow diagrams (Figures S2, S3, S4, and S5).

Extraction and Effect Size Calculation

Once studies had been selected for inclusion, we extracted the information necessary for the analysis. For data extraction, two individual study members extracted the data from each study to ensure the accuracy of the extracted information. We extracted information either directly from the articles, from figures using the metaDigitise package in R (Pick et al., 2018), or when available, the raw study data. Studies that provided either insufficient information or overpopulated figures from which it was not possible to extract reasonably accurate outcome values or approximate sample sizes were excluded from the analyses.

Because the included studies quantified the association between age and economic preferences using different metrics and study designs, before combining all the outcomes in the meta-analysis, we first converted these into correlation coefficients.

For studies using an extreme group design where the outcome variable was continuous but age was dichotomous (i.e., younger and older adults), we converted the standardized mean difference (or t-test value) between two age groups into a point-biserial correlation coefficient. However, if Pearson's r correlation coefficients between age and outcomes had been or could be calculated, these were selected for the analysis. For designs where both age and the outcome variable were measured continuously, we used Pearson's r correlation coefficient.

We coded all effect sizes such that higher values indicated either increasing risk taking, altruism, temporal discounting, or effort discounting with age. For extreme group designs, we focused on the differences between the youngest and oldest adult samples, and did not include in the analyses differences with intermediate age groups.

By following this procedure, we created four sets of effect sizes (i.e., one for each preference), which resulted in a total of 369 effect sizes with data from 141,794 individuals. In addition, to subsequently assess the effect of certain moderators on the individual effect sizes, we coded: (a) the type of study design from which it originated (i.e., extreme design or continuous), (b) the effect size metric (i.e., Pearson's r or point-biserial correlation), (c) whether the task involved hypothetical or incentivized decisions, (d) decisions from experience or description, (e) whether these decisions were made in the gain or loss domain, (f) study context (i.e., online or in person), and (g) proportion of females in the sample (see Table S1 for rationale). In addition, we calculated the age range of the sample. For Pearson's r correlations, we computed the age difference (in decades) between the youngest and oldest participant, and for point-biserial correlations, the difference between the mean age (in decades) of the oldest and youngest adult group. If this information was missing, we used the midpoint of the age range of each group (e.g., if participants in a group were between 18 and 30 years of age, we used 24 as the value).

A detailed overview of the included studies, as well as the data and code used to compute effect sizes, is available in the online repository.

Analysis

We used the R statistical environment (R Core Team, 2020) and the metafor package (Viechtbauer, 2010) to perform the analyses. The analysis code is available in the online repository.

Meta-Analysis

Some studies reported multiple outcomes (e.g., multiple conditions, multiple behavioral indices); instead of selecting one outcome per study or aggregating these, we entered all outcomes in the meta-analysis. For each data set we fitted a three-level meta-analystic model using restricted maximum likelihood (REML) estimation. The model

included random effects at the estimate (i.e., level 2 cluster variable) and study (i.e., level 3 cluster variable) levels, and accounted for the dependence of effect sizes by allowing the sampling errors within studies to be correlated. A correlation of 0 would indicate that the outcomes are independent whereas a correlation of 1 would indicate full correspondence; for our analyses, we opted for a correlation of 0.5. To explore whether the level of correlation between outcomes of the same study had an influence on the results, we ran sensitivity analyses with correlations varying between .1 and .9 (Figure S12). Additionally, we applied robust variance estimation methods to obtain more precise model estimates (Pustejovsky & Tipton, 2022).

In addition to assessing the statistical significance (alpha = 5%) of the meta-analytic effect size estimates, we assessed their practical significance by performing equivalence tests (Lakens, 2017; Lakens et al., 2018). Based on standard guidelines (Cohen, 1988), we chose $r = |.1|$ as the smallest effect size of interest; this is defined as a small effect, but representative of the correlations found in individual differences research (Bosco et al., 2015; Gignac & Szodorai, 2016).

To identify whether any study was particularly influential, we conducted on each set of effect sizes an influence analysis by computing the pooled effect size omitting one study at a time (Figure S11).

Lastly, informed by previous meta-analyses and theory, we estimated a series of meta-regression models to test whether some of the heterogeneity in age-related effects could be explained by certain moderators (see Table S1 for an overview).

Cumulative Meta-Analysis and Historical Trends

Cumulative meta-analyses can be conducted by adding effect sizes to the meta-analytic model in chronological order by study or by publication year. With the latter approach, we can better examine temporal trends while also accounting for cases when more

than one study can be published within the same year, which, depending on how they are entered in the cumulative meta-analysis, could affect the shape of the plots (Koricheva et al., 2013; Leimu & Koricheva, 2004). Therefore, we prioritize reporting the results from repetitively fitting the above-specified three-level meta-analysis model by adding the effect sizes by publication year. The results of the cumulative meta-analyses conducted at the study level are reported in the Supplementary Appendix.

To explore historical trends in effect sizes, we included in a meta-regression model the number of decades the study had been published as of 2022 as a moderator. Additionally, we explored changes in sample sizes (log) over time by fitting a linear regression. Further, considering the predictions made by certain theories (e.g., socio-emotional selectivity theory, confound hypothesis), for risk preference, we explored the accumulation of evidence and temporal trends for each domain (i.e., gain, loss and mixed) and task type (i.e., description and experience) separately. This amounts to conducting an independent estimation of residual between-studies variance for the two moderators (cf., Rubio-Aparicio et al., 2020). Lastly, we explored the relation between yearly citations with effect sizes and sample sizes. Details on the method used and results are available in the Supplementary Appendix.

Publication Bias

For all four sets of effect sizes, we performed various analyses, including Egger's tests and p-curve tests, and produced funnel plots to check for publication bias in the published literature (see the Supplementary Appendix for details on our approach and results).

Results

There were differences across the four economic preferences in the number of effect sizes, their distribution, and study sample size (Figure S6). Here, we report on the overall effect of age based on cluster-robust inference and the accumulation of evidence for each

economic preference separately by displaying effect sizes by year of publication (study-level estimates are included in the Supplementary Appendix; Figure S7).

Risk

Meta-Analysis

The meta-analysis of the relevant 193 effect sizes suggest age is not associated with risk preference (r = -0.02, 95% CI [-0.06, 0.02], p = 0.251). Equivalence tests showed that the effect fell within the equivalence bounds (z = 3.73, p < 0.001; Figure 2A).

To investigate effect-size heterogeneity, Q(df = 192) = 2232.19, p < 0.001, we estimated in separate three-level meta-regressions the moderating role of (a) sample age range, (b) gender, (c) effect size metric, (d) study design, (e) incentivization, (f) domain, (g) task type, and (h) study context. We find a small but significant decrease in risk-taking with age in the gain domain, none of the other moderators were statistically significant (Table S4). Further, when we performed separate analyses for each domain (Figure S8), allowing the amount of residual heterogeneity to differ between domains (versus using a pooled estimate as with the meta-regression), this effect remained (r = -0.06, 95% CI [-0.11, -0.02], p = 0.005), with an equivalence test showing that this effect fell outside the equivalence bounds (Figure 2B). Further, for decisions from experience, although in the separate analyses the effect remained statistically non-significant, we cannot reject that the association between risk taking in these tasks and age is at least -0.1 (Figure 2C).

Cumulative Meta-Analysis and Historical Trends

From Figure 1C, we observe that there was never any evidence supporting age differences in overall task-based risk-taking. Since 2010, effect sizes have remained relatively stable, and oscillated between -0.07 and -0.01. Further, when splitting effect sizes by domain or task, as shown in Figure S8C and Figure S9C, the effect of age is not particularly stable over time and the number of effect sizes in each category is quite

heterogeneous, thus warranting additional evidence in each domain and type of task. There was no linear relation between effect size (b = -0.03, p = 0.434), or sample size and number of decades the paper had been published for (Figure S10). Further, we find a statistically significant linear effect of sample size on citations but not of effect size (Figure S15, Tables S10 and S11).

Summary

To summarize, we find overall no effect of age on risk preference. Concerning moderators, and contrary to previous syntheses, we find no strong support for the idea that age differences vary systematically as a function of the learning and memory demands of the task as captured through the distinction between description versus experience (Mata et al., 2011). However, we find a small negative effect of age in the gain domain in line with past meta-analytic work (Best & Charness, 2015) and some theoretical predictions (Depping & Freund, 2011). Concerning historical trends, we find no evidence of trends in effect sizes, samples sizes, or citations. Finally, we find overall no evidence of publication bias (see Supplementary Appendix).

Time

Meta-Analysis

The meta-analysis of the 125 effect sizes shows a small negative effect of age on time preference (r = -0.04, 95% CI [-0.07, -0.01], p = 0.020). However, equivalence tests showed that the effect fell within the equivalence bounds (Figure 2A).

To understand possible differences between the individual effect sizes (Q(df = 124) = 496.69, p < 0.001), we conducted separate meta-regressions to investigate the moderating role of (a) sample age range, (b) gender, (c) effect size metric, (d) study design, (e) incentivization, and (f) study context. We noted a significant difference in effects for study context: we find an age-related effect for temporal discounting in online studies (b = -0.05,

95% CI [-0.08, -0.02], $p = 0.002$). Further, we also note effects of study design and effect size metric (Table S5). Lastly, there is also a difference due to incentives, but given that close to 90% of the studies included in our analyses involve hypothetical payments, we treat this difference with caution.

Cumulative Meta-Analysis and Historical Trends

Figure 1C shows that the first study published in 1994 (i.e., Green et al., 1994), found a large age difference, with older adults exhibiting less temporal discounting than younger adults ($r = -0.72$ SE = 0.22). However, in 2002, the second study was published (i.e., Kirby et al., 2002) reporting evidence in the opposite direction ($r = 0.30$, SE = 0.07), and when combining this with the evidence from the first study, it led the pooled effect size to shift closer to zero, increased the uncertainty around it, and made it statistically non-significant ($r = -0.19$, 95% CI [-6.66, 6.28], $p = 0.774$). Since then, age differences in temporal discounting have remained non-significant, with pooled effect sizes nearing zero, but more recently such small negative effects reached statistical significance. We tested for the presence of the Proteus phenomenon, which is when a large and extreme result is first published but is followed by the publication of less extreme results and can be indicative of publication bias (Ioannidis & Trikalinos, 2005; Young et al., 2008). We followed the approach by Koricheva et al. (2013) and compared the effect size and variance of the first study with the mean effect size and variance of the rest of the published studies. We obtained a z-value of 2.71, $p = 0.007$, suggesting that the study by Green et al. (1994) differed significantly above chance from the other results.

The historical analyses showed no linear effect of decades since the paper has been published on the size of the effects ($b = 0.04$, $p = 0.538$). Furthermore, there was no statistically significant linear relation between publication year and sample size (Figure S10). Concerning the citation analyses, we do not detect any discernible trend (Figure S15).

Summary

We observe a small negative effect of age on time preference, however, equivalence tests show that this effect can be considered trivially small. Regarding historical trends, we find no evidence of trends in effects sizes, sample sizes, or citation patterns. Concerning publication bias, we find evidence of a Proteus phenomenon but no other evidence of bias (see Supplementary Appendix).

Social

Meta-Analysis

The meta-analysis of the 28 effect sizes revealed a small positive effect of age on social preference, suggesting that altruistic behavior as measured by behavioral tasks increases with age (r = 0.11, 95% CI [0.01, 0.21], p = 0.033). This is consistent with the results from the recent meta-analysis by Sparrow et al. (2021), who also reported a positive, albeit larger, effect size (r = 0.24, 95% CI [0.12, 0.35], p = 0.001)[1]. Further, this effect also falls outside the equivalence bound, but is not distinguishable from the upper bound (Figure 2A).

As there was considerable heterogeneity in the effect sizes (Q(df = 27) = 265.65, p < 0.001), we also explored the potential moderating role of (a) sample age range, (b) gender, (c) effect size metric, (d) study design, (e) incentivization, and (f) study context. We find that this positive age effect is mainly driven by point-biserial correlation coefficients (k = 14; b = 0.17, 95% CI [0.05, 0.29], p = 0.011). Out of the rest of the moderators, we also noted an effect of study design (Table S6).

Cumulative Meta-Analysis and Historical Trends

Relative to age differences in risk or time preference, age differences in social preference have been more recently investigated (Figure 1). Initially, no significant age differences were reported; however, with additional studies reporting larger (and statistically

significant) effect size estimates, the cumulative estimate began to shift away from zero in the positive direction. It reached a peak (r = 0.16, SE = 0.05) in the year 2018 (includes 10 studies and 12 effect sizes); however, since then, effect size estimates published were zero (Figure S7), moving the pooled effect size closer to the null. We find no statistically significant linear effect of decades since publishing on effect sizes ($b = 0.10$, $p = 0.448$), showing that over the years the effect sizes have remained generally comparable. Although we visually note an increase in study sample sizes over the years, it was not statistically significant (Figure S10). We find no evidence for trends in citation patterns, except for studies with smaller samples getting more cited (Figure S15, Table S11).

Summary

We find an overall positive effect of age on social preference but this effect is smaller than previous published estimates ($r = 0.11$ vs. 0.24; Sparrow et al., 2021). Concerning moderators, we find some evidence for an effect of effect size metric and study design. We find little evidence of temporal trends. Concerning publication bias, additional analyses using Egger's regression provide some evidence of publication bias (see Supplementary Appendix).

Effort

Meta-Analysis

The meta-analysis of 23 effect sizes revealed a positive but not significant effect of age on effort discounting ($r = 0.24$, 95% CI [-0.05, 0.52], $p = 0.087$). Further, from the equivalence tests, we note that the upper bound equivalence test was non-significant (Figure 2A); therefore, we cannot reject that the association between effort discounting and age is different from 0.1.

We observe substantial heterogeneity, $Q(\text{df} = 22) = 132.60$, $p < 0.001$), despite the small number of studies included (s = 7). We explored the potential moderating role of of (a) sample age range, (b) gender, (c) effect size metric, (d) effort type, and (e) domain (Table

S7). We did not consider incentivization, study context, or study design as moderators because all studies were conducted in a laboratory context, and except for one study, involved incentivized decisions and had an extreme group design. Out of the included moderators, effort type had statistically significant effect on the observed outcomes. Cognitive effort discounting was greater for older than younger adults (b = 0.47, 95% CI [0.39, 0.55], p = 0.001).

Cumulative Meta-Analysis and Historical Trends

Similar to age differences in social preference, age differences in effort discounting have been more recently investigated (Figure 1). Initially, a first article (Westbrook et al., 2013) was published reporting significant age differences (r = 0.53 SE = 0.07), but subsequent studies provided mixed results. We tested for the presence of the Proteus Phenomenon, and obtained a z-value of -0.89, p = .375, suggesting that the results by Westbrook et al. (2013) did not differ significantly above chance from the other results. Given the small sample size of these studies, error is wide (Figure S7) and the pooled effect size has a quite wide error range. Within the brief time that age differences in effort discounting have been investigated, we find no statistically significant linear effect of decades since publishing on effect sizes (b = 0.32, p = 0.333), nor an increase in study sample sizes over the years (Figure S10).

Summary

We find an overall positive but not significant effect of age on effort preferences. Concerning moderators, there is evidence for the role of effort type (i.e., physical vs. cognitive) suggesting that there is an effect of age on effort discounting specific to cognitive effort. Yet, given the small number of studies included in our analysis, further evidence for both types of effort is still required to assess the robustness of this result. Concerning

temporal trends, we find no discernible trends in effect sizes or sample sizes. Finally, additional analyses show no evidence of publication bias (see Supplementary Appendix).

Discussion

We aimed to contribute to a better understanding of the match between extant theoretical accounts of age differences in economic preferences and the associated empirical literature by providing a tabular overview of theories that have been used to make predictions about age differences in economic preferences and conducting a quantitative synthesis of the results of behavioral studies. For this purpose, we conducted systematic literature searches and meta-analyses to estimate overall age effects in risk, time, social, and effort preferences. We also investigated the role of possible moderators, including domain (e.g., gain vs. loss), measurement characteristics (e.g., description vs. experience, incentivization), and study or sample characteristics (e.g., proportion females). Furthermore, we assessed historical trends in evidence accumulation through the use of cumulative meta-analysis and by exploring historical trends in research practices (e.g., sample sizes). All in all, we hoped our approach could provide an assessment of the adequacy of different theories of age differences in economic preferences to account for the current and past empirical record.

Main findings

Overall, our meta-analyses identified non-significant effects of age for risk (r = -0.02, 95% CI[-0.06, 0.02]), and effort (r = 0.24, 95% CI[-0.05, 0.52]) preferences, and a small but significant effect of age for social (r = 0.11, 95% CI[0.01, 0.21]) and time (r = -0.04, 95% CI[-0.07, -0.01]) preferences, suggesting increased altruism and patience with age, respectively. More generally, we find all effects are small and cannot be fully distinguished from an equivalence bound of r = |0.1|, which can be considered a practically or theoretically meaningful interval.

Taken together, these results suggest either non-existent or small effects of age in economic preferences. These results are compatible with past meta-analytic work on risk (Best & Charness, 2015; Mata et al., 2011), which did not show an overall effect of age on risk taking in behavioural tasks. For time, our results are similar to those of a previous meta-analysis (Seaman et al., 2022) that reported a small negative, albeit non-significant effect of age on temporal discounting. In turn, the results for social preferences are smaller in magnitude than the previous meta-analytic estimate (Sparrow et al., 2021). Finally, the meta-analytic result for effort preferences reflects the mixed findings observed in primary studies of age differences in this area (e.g., Hess et al., 2021; Westbrook et al., 2013).

Concerning the analysis of moderators, our results are particularly noteworthy in the context of risk preferences for which different theories have been proposed that make specific predictions about different moderators. In line with past syntheses (Best & Charness, 2015) and theories that foresee differential age effects as a function of gain and loss domains (cf. Depping & Freund, 2011), we find evidence of age differences in risk preference in the gain relative to the loss domain. Furthermore, contrary to predictions from the confound hypothesis (Frey et al., 2021; Olschewski et al., 2018) and past empirical results (Mata et al., 2011), we do not find a significant pattern of larger age effects in decisions from experience. The main reason for these differences appears to be the inclusion of novel evidence relative to the previous meta-analysis (Mata et al., 2011). Overall, the role of other moderators, such as the use of incentivization, does not seem to account for systematic variance in effect sizes in economic preferences, but some methodological choices (i.e., correlation type, study design) do account for some variance in the social and time preference domain. Furthermore, for temporal discounting, we observe an age difference in online relative to laboratory studies: Laboratory and online studies may differ in their sample characteristics and it would

be interesting to assess the extent to which sample composition (e.g., education level) accounts for such differences in future work.

Concerning historical trends, the apparent visual trend across economic preferences is for effect sizes to approach zero over time; however, we found overall no evidence of significant effects over time for either effect sizes or research practices as quantified by sample size of the studies conducted. As noted in earlier work (Seaman et al., 2022), the results for time preference make clear that the overall null effect of age on temporal discounting was already apparent early in the research history of the topic, because the large effect reported in the seminal paper was not replicated in subsequent studies (Green et al., 1994). More broadly, one should note that the four types of economic preferences differ considerably in the number of effect sizes available for analysis (193, 125, 28, 23, for risk, time, social, and effort preferences, respectively), suggesting it could be important to assess the development of such trends in future work, particularly for the social and effort preferences for which comparatively little evidence is available.

Finally, concerning our analyses of publication bias, p-curve analyses found no evidence of p-hacking but we found evidence of a Proteus effect (i.e., the tendency for early replications of a scientific work to contradict the original findings; Ioannidis & Trikalinos, 2005; Koricheva et al., 2013; Young et al., 2008) in the time preferences literature and Egger's regression provided some ground to suspect systematic publication bias in the social preferences literature. These results do not fully assuage concerns surrounding the overestimation of age effects in the aging literature (Isaacowitz, 2020), but also do not provide evidence for widespread publication bias.

Implications

All in all, our results have some major theoretical and methodological implications. First and foremost, concerning theory, our finding of small to null age effects detected across

the empirical literature questions the adequacy of many extant theories that predict age differences in economic preferences. One direct consequence is that the theoretical perspectives concerning risk preferences need to be revised. Indeed, our results reject theories that posit a strong role for cognitive and learning effects (cf. Mata et al., 2011), but provide support for theories predicting differential age effects as a function of gain and loss domains (cf. Depping & Freund, 2011). We propose that future theorizing should focus more specifically on the mechanisms thought to underlie age differences (e.g., dopaminergic function, time horizon) and empirical work should aim to provide critical tests of the role of such mechanisms (cf. Frey et al., 2015; Zilker & Pachur, 2021) rather than simply assess a directional effect of age. It may also be important to distinguish critical claims of theories, such as the age trends associated with specific mechanisms, and auxiliary assumptions, such as the role of task or measurement characteristics (e.g., role of incentivization, task complexity). We discuss the specific point concerning assumptions about operationalization in the Limitations section below.

Second, concerning methodological implications, the few indications of publication bias suggest future work may want to consider different sources of bias and the use of registered reports to correct our estimates of age differences in economic preferences.

Third, and more broadly, even though we could not distinguish clear-cut phases in the development of the research topic, we would like to encourage researchers studying aging to integrate cumulative approaches in their work. Here, we focused on economic preferences but this approach could be extended to other central constructs in aging research, such as memory performance, executive functioning, or well-being. In doing so, we could detect areas in which age differences are more established, robust, and stable than others, which, ultimately, could improve how we justify the need for additional research, how resources are allocated, and how participants are recruited.

To summarise, our meta-analysis did not find evidence to support the predictions made by the theories that are most frequently discussed in the literature on aging and preferences for risk and effort. For time preference, more than half of these theories (e.g., dopaminergic neuromodulation hypothesis) predict a decrease in temporal discounting with age, however given that effect we identified is of very small magnitude, the extent to which these theories are supported is questionable. When it comes to social preference, our results suggest that there is a small increase in altruism with age, which is consistent with the predictions made by close to all the theories that we examined in this domain. However, there are relatively few studies concerning social and effort preferences, and our results do not provide sufficient evidence to distinguish between the various mechanisms proposed suggesting more work is needed in the area of economic preferences.

Limitations and Future Directions

We should also point out some limitations of our work. First, a wide range of measures has been developed to quantify individuals' economic preferences (Charness et al., 2013; Eckel, 2019). In the present study, we focus solely on behavioral tasks, yet self-reported measures (e.g., propensity measures) could also be considered. The convergent validity of different measures within each preference is low (Duckworth & Kern, 2011; Frey et al., 2017; Levitt & List, 2007; Strand et al., 2018), which suggests further research should focus on the comparability of effect size trajectories across different measurement types. For example, recent work suggests that self-reports are more likely to capture systematic age differences in risk preference (Frey et al., 2021) and a recent quantitative synthesis suggests robust age effects when considering self-report measures (Liu et al., 2023). Although past theorizing has largely ignored the role of measurement, the differences between our results and those for self-reported risk propensity (Liu et al., 2023) suggest that it would be

important to develop more specific expectations about the role of operationalization in detecting age differences in economic preferences.

Second, our work focused solely on published results because of our aim of assessing the historical patterns in the literature. However, published results are unlikely to be fully representative of the evidence on age differences thus data from unpublished reports or data sets could be included in future extensions of this work.

Third, although we considered a wide range of moderators to explain effect size heterogeneity, cultural and socio-demographic factors (e.g., education) were not included. Details on such factors are often missing in primary studies or reported heterogeneously, which can be challenging to incorporate in analyses. However, as such factors can influence economic preferences (cf. Frey et al., 2021), this can be an avenue for future research.

Lastly, we did not preregister this work. We note, however, that we make all the data and code used in this study publicly available to ensure that our work can be assessed transparently and used in future confirmatory efforts.

Conclusion

Our results indicate that age differences in economic preferences as captured by behavioral tasks are not as pervasive as extant theories would imply, and that more specific theorizing is needed to make predictions for different preference types (risk, time, social, effort) and their operationalizations.

Footnotes

[1]Sparrow et al. (2021) reported an overall effect of g = .61 (r = .31; 95%CI[0.25, 0.37]; p <.001), however, one of the outcomes used in their analyses was coded in the opposite, incorrect direction (https://osf.io/9hacs). Upon correction, the mean effect size becomes g =.48 which we converted into a correlation coefficient (r = .24; 95%CI[0.12, 0.35]; p = .001).

Acknowledgments

The authors thank Laura Wiles for editing the manuscript.

This study was not preregistered, but all the data and scripts are publicly available in an online repository (https://github.com/cdsbasel/cumulative).

Funding

This work was supported by grants from the Swiss National Science Foundation to R.Mata (https://data.snf.ch/grants/grant/204700, https://data.snf.ch/grants/grant/177277).

Author contributions

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

Y.L.: Investigation and Validation.

A. Bajrami: Investigation and Validation.

G.S.: Investigation and Validation.

L.T.: Investigation, Methodology, Validation, and Writing - review & editing.

R.M.: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, and Writing - review & editing.

Competing interests

All authors declare no competing interests related to this study

References

Best, R., & Charness, N. (2015). Age differences in the effect of framing on risky choice: A meta-analysis. *Psychology and Aging, 30* (3), 688–698. https://doi.org/10.1037/a0039447

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*, 431–449. https://doi.org/10.1037/a0038047

Carstensen, L. L. (2006). The influence of a sense of time on human development. *Science, 312*(5782), 1913-1915. https://doi.org/10.1126/science.1127488

Charness, G., Gneezy, U., & Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization, 87*, 43–51. https://doi.org/10.1016/j.jebo.2012.12.023

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillside, NJ: Lawrence Erlbaum Associates.

Depping, M. K., & Freund, A. M. (2011). Normal aging and decision making: The role of motivation. *Human Development, 54* (6), 349–367. https://doi.org/10.1159/000334396

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

Eckel, C. C. (2019). Measuring individual risk preferences. *IZA World of Labor.* https://doi.org/10.15185/izawol.454

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior, 6* (3), 347–369. https://doi.org/10.1006/game.1994.1021

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature, 40* (2), 351–401. https://doi.org/10.1257/002205102320161311

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

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances, 3* (10), e1701381. https://doi.org/10.1126/sciadv.1701381

Frey, R., Richter, D., Schupp, J., Hertwig, R., & Mata, R. (2021). Identifying robust correlates of risk preference: A systematic approach using specification curve analysis. *Journal of Personality and Social Psychology, 120* (2), 538–557. https://doi.org/10.1037/pspp0000287

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Green, L., Fry, A. F., & Myerson, J. (1994). Discounting of delayed rewards: A life-span comparison. *Psychological Science, 5* (1), 33–36. https://doi.org/10.1111/j.1467-9280.1994.tb00610.x

Hess, T. M., Lothary, A. F., O'Brien, E. L., Growney, C. M., & DeLaRosa, J. (2021). Predictors of engagement in young and older adults: The role of specific activity experience. *Psychology and Aging, 36* (2), 131–142. https://doi.org/10.1037/pag0000561

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

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology, 58* (6), 543–549. https://doi.org/10.1016/j.jclinepi.2004.10.019

Isaacowitz, D. M. (2018). Planning for the future of psychological research on aging. *The Journals of Gerontology: Series B, 73* (3), 361–362. https://doi.org/10/ggb78h

Isaacowitz, D. M. (2020). Doing more with null age effects: Introduction to the special section. *The Journals of Gerontology: Series B, 75* (1), 42–44. https://doi.org/10/gnnzcq

Kirby, K. N., Godoy, R., Reyes-Garcı´a, V., Byron, E., Apaza, L., Leonard, W., Pérez, E., Vadez, V., & Wilkie, D. (2002). Correlates of delay-discount rates: Evidence from Tsimane' Amerindians of the Bolivian rain forest. *Journal of Economic Psychology, 23* (3), 291–316. https://doi.org/10.1016/S0167-4870(02)00078-8

Koricheva, J., Jennions, M. D., & Lau, J. (2013). Temporal trends in effect sizes: Causes, detection, and implications. In J. Koricheva, J. Gurevitch, & K. Mengerson (Eds.), Handbook of meta-analysis in ecology and evolution (pp. 237–254). Princeton University Press.

Kulinskaya, E., & Mah, E. Y. (2022). Cumulative meta-analysis: What works. *Research Synthesis Methods, 13* (1), 48–67. https://doi.org/10.1002/jrsm.1522

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science, 8* (4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science, 1* (2), 259–269. https://doi.org/10.1177/2515245918770963

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). Meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine, 327* (4), 248–254. https://doi.org/10/c4wnw4

Leimu, R., & Koricheva, J. (2004). Cumulative meta–analysis: A new tool for detection of temporal trends and publication bias in ecology. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 271* (1551), 1961–1966. https://doi.org/10/bv6kfh

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives, 21* (2), 153–174. https://doi.org/10.1257/jep.21.2.153

Liu, Y., Bagaïni, A., Son, G., Kapoor, M., & Mata, R. (2023). Life-course trajectories of risk-taking propensity: A coordinated analysis of longitudinal studies. *The Journals of Gerontology: Series B, 78*(3), 445-455. doi: 10.1093/geronb/gbac175

Löckenhoff, C. E., & Rutt, J. L. (2017). Age differences in self-continuity: Converging evidence and directions for future research. *The Gerontologist, 57*(3), 396-408.

Mata, R., Josef, A. K., Samanez-Larkin, G. R., & Hertwig, R. (2011). Age differences in risky choice: A meta-analysis. *Annals of the New York Academy of Sciences, 1235* (1), 18–29. https://doi.org/10.1111/j.1749-6632.2011.06200.x

Mayr, U., & Freund, A. M. (2020). Do we become more prosocial as we age, and if so, why? *Current Directions in Psychological Science, 29,* 248–254. https://doi.org/10.1177/0963721420910811

Olschewski, S., Rieskamp, J., & Scheibehenne, B. (2018). Taxing cognitive capacities reduces choice consistency rather than preference: A model-based test. Journal of experimental psychology. *General, 147*, 462–484. https://doi.org/10.1037/xge0000403

Ostaszewski, P., Bąbel, P., & Swebodziński, B. (2013). Physical and cognitive effort discounting of hypothetical monetary rewards. *Japanese Psychological Research, 55* (4), 329–337. https://doi.org/10.1111/jpr.12019

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *British Medical Journal, 372*, n71. https://doi.org/10.1136/bmj.n71

Pick, J., Nakagawa, S., & Noble, D. (2018). Reproducible, flexible and high-throughput data extraction from primary literature: The metaDigitise R package. *Biorxiv*. https://doi.org/10.1101/247775

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science, 23* (3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rubio-Aparicio, M., López-López, J. A., Viechtbauer, W., Marín-Martínez, F., Botella, J., & Sánchez-Meca, J. (2020). Testing categorical moderators in mixed-effects meta-analysis in the presence of heteroscedasticity. *The Journal of Experimental Education, 88*(2), 288-310. https://doi.org/10.1080/00220973.2018.1561404

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studiesures, and More Online Studies. *Advances in Methods and Practices in Psychological Science, 2* (2), 107–114. https://doi.org/10.1177/2515245919838781

Seaman, K. L., Abiodun, S. J., Fenn, Z., Samanez-Larkin, G. R., & Mata, R. (2022). Temporal discounting across adulthood: A systematic review and meta-analysis. *Psychology and Aging, 37* (1), 111. https://doi.org/10.1037/pag0000634

Seaman, K. L., Gorlick, M. A., Vekaria, K. M., Hsu, M., Zald, D. H., & Samanez-Larkin, G. R. (2016). Adult age differences in decision making across domains: Increased discounting of social and health-related rewards. *Psychology and Aging, 31* (7), 737–746. https://doi.org/10.1037/pag0000131

Sparrow, E. P., Swirsky, L. T., Kudus, F., & Spaniol, J. (2021). Aging and altruism: A meta-analysis. *Psychology and Aging, 36* (1), 49–56. https://doi.org/10/gj2gsj

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring Listening Effort: Convergent Validity, Sensitivity, and Links With Cognitive and Personality Measures. *Journal of speech, language, and hearing research: JSLHR, 61* (6), 1463–1486. https://doi.org/10.1044/2018_JSLHR-H-17-0257

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1–48. https://doi.org/10.18637/jss.v036.i03

Westbrook, A., Kester, D., & Braver, T. (2013). What Is the Subjective Cost of Cognitive Effort? Load, Trait, and Aging Effects Revealed by Economic Preference. *PloS one, 8,* e68210. https://doi.org/10.1371/journal.pone.0068210

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why Current Publication Practices May Distort Science. *PLOS Medicine, 5* (10), e201. https://doi.org/10.1371/journal.pmed.0050201

Zilker, V., & Pachur, T. (2021). Does option complexity contribute to the framing effect, loss aversion, and delay discounting in younger and older adults? *Journal of Behavioral Decision Making, 34* (4), 488–503. https://doi.org/10.1002/bdm.2224

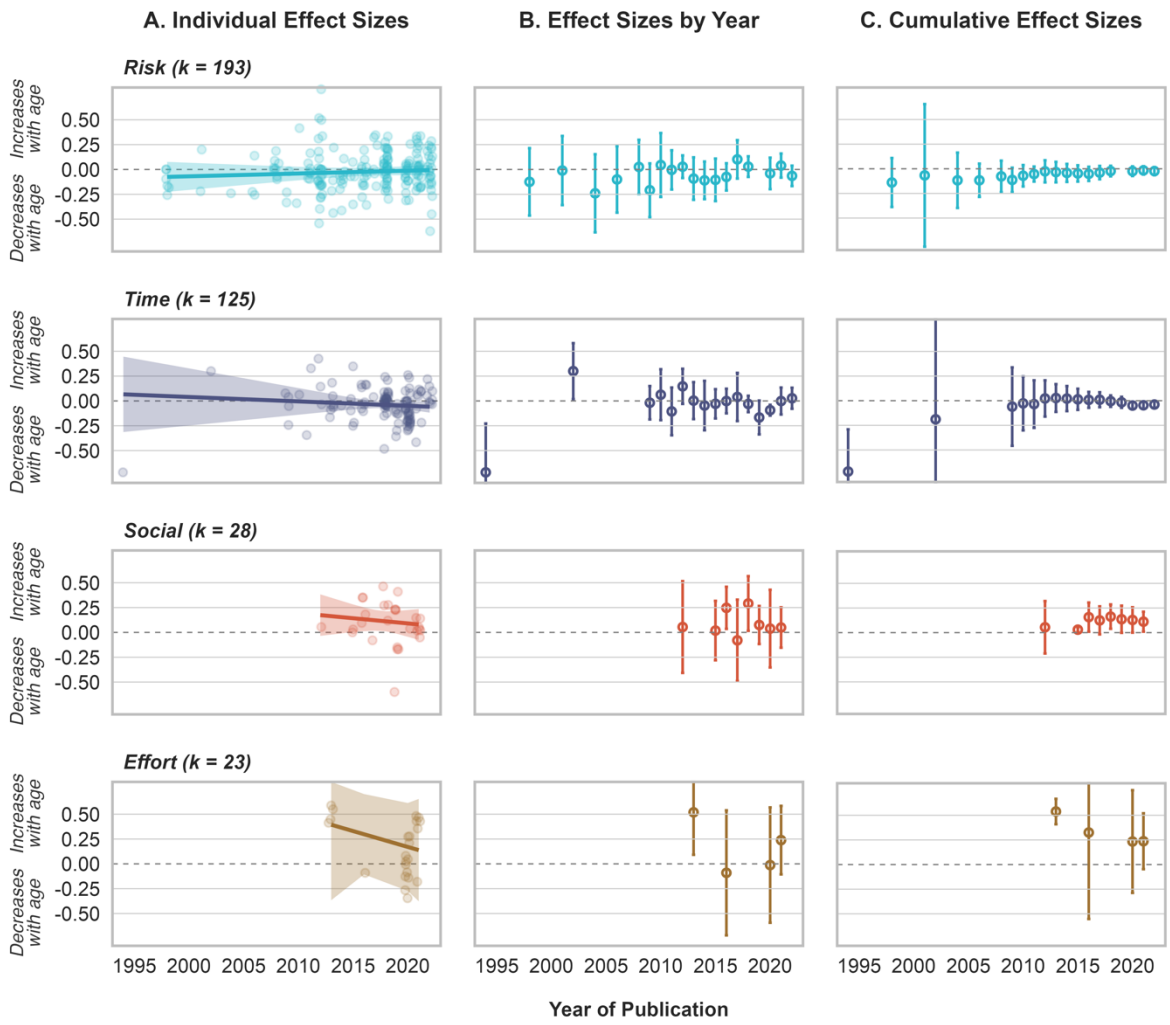


Figure 1. Meta-analytic results of the effect of age on risk (k = 193, s = 62, n = 39832), time (k = 125, s = 54, n = 115496), social (k = 28, s = 15, n = 2997), and effort (k = 23, s = 7, n = 571) preferences. (A) Scatter plots of the individual effect sizes plotted as a function of the publication year with model predictions and 95% CI. (B) Aggregated forest plots of the three-level meta-analytic model with effect sizes pooled by year with 95% CI. (C) Forest plots of the cumulative effect sizes and 95% CI based on cluster-robust inference. CI = confidence interval.

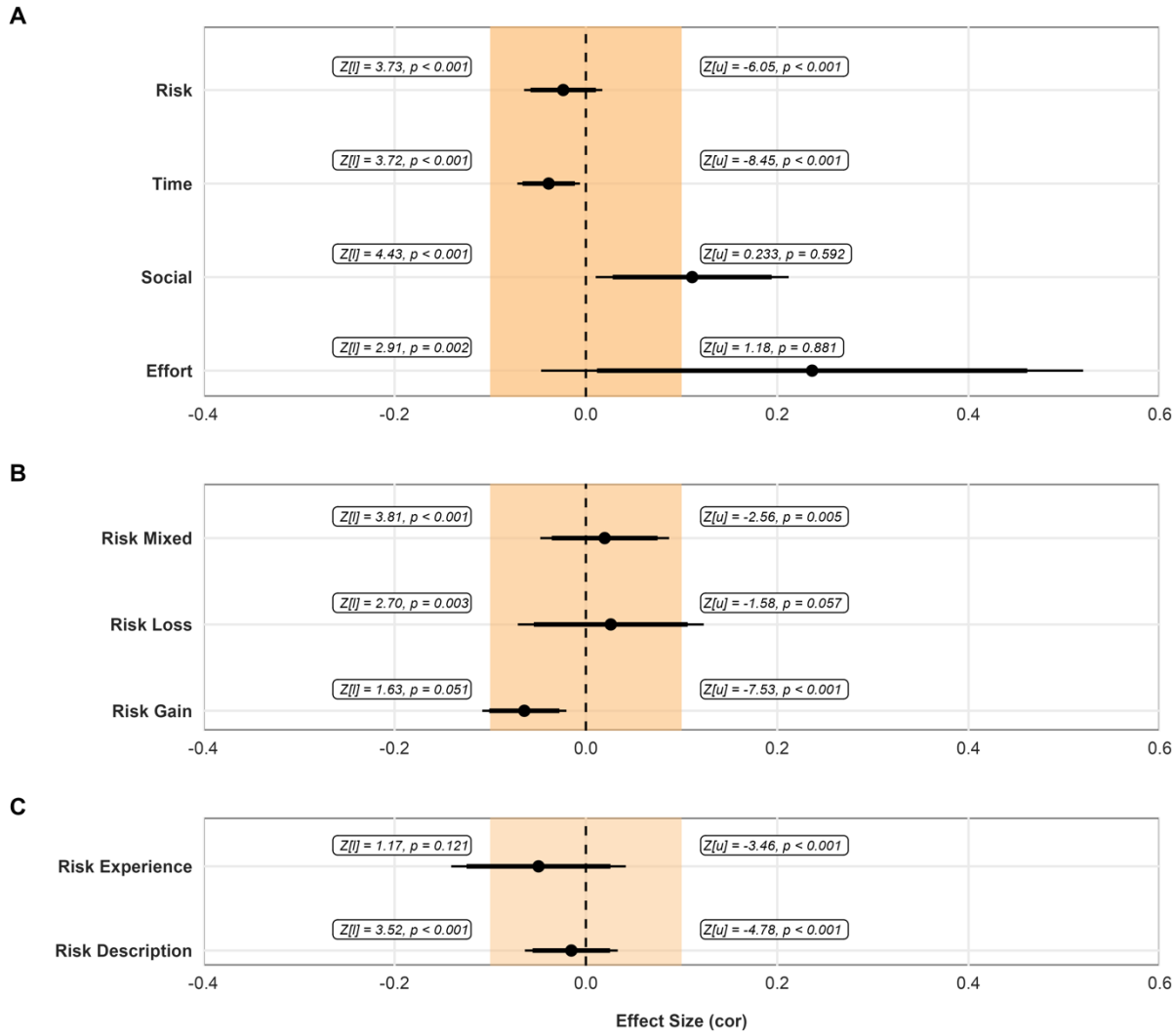


Figure 2. Equivalence test results (against the [u]pper and [l]ower equivalence bounds) for the estimated pooled effect sizes (dots), with 95% (thin lines) and 90% (thick lines) cluster-robust confidence intervals. The shaded section represents the equivalence bounds (r = |0.1|). (A) Pooled effect size estimates from the separate three-level meta-analysis models for age differences in risk (k = 193), time (k = 125), social (k = 28), and effort (k = 23) preference. (B) Pooled effect size estimates from the separate three-level meta-analysis models for age differences in risk taking in the gain (k = 106), loss (k = 46), and mixed (k = 41) domain. (C) Pooled effect size estimates from the separate three-level meta-analysis models for age differences in risk taking in decisions from description (k = 147) and experience (k = 44).

Table 1

Description of Economic Preferences.

Preference	Description	Trade-Off	Examples of Measures
Risk	Propensity to choose options with higher variance in potential monetary outcomes	smaller-sure vs. larger-uncertain	Cups Task, Lotteries, Bal- loon Analogue Risk Task
Time	Valuation placed on receiving a monetary reward sooner compared to later	smaller-sooner vs. larger-later	Temporal Discounting Task
Social	Propensity of an individual to engage in behavior motivated by the concern for others and for the sake of others	smaller-self vs. larger-other	Dictator Game, Social Dis-counting Task
Effort	Valuation placed on a monetary reward after exerting physical or cognitive effort.	Low reward-low effort vs. large reward-large effort	Effort Discounting Task

Table 2

Theoretical Accounts of Associations Between Age and Economic Preferences. Adapted from Frey et al. (2021) and Seaman et al. (2022)

Theoretical Account	Preference			
	Risk	Time	Social	Effort
Affective forecasting (cf. Seaman et al., 2022)	.	D	.	.
Cognitive control hypothesis (cf. Frey et al., 2021)	D	.	.	.
Confound hypothesis (cf. Frey et al., 2021; Mata et al., 2011; Seaman et al., 2022)	I/D	I/D	.	.
Dopaminergic neuromodulation hypothesis (cf. Frey et al., 2021; Seaman et al., 2022; Westbrook et al., 2013)	D	D	.	I
Intuitive-prosociality hypothesis (cf. Mayr & Freund, 2020)	.	.	I	.
Evolutionary signaling hypothesis (cf. Frey et al., 2021)	D	.	.	.
Future self-continuity hypothesis (cf. Löckenhoff, & Rutt, 2017)	.	D	.	.
Mortality (cf. Seaman et al., 2022)	.	U	.	.
Motivational re-orientation theory (c.f. Frey et al., 2021; Hess et al., 2021)	I/D	.	.	I/D
Psychosocial Development theory (cf. Mayr & Freund, 2020)	.	.	∩	.
Resource accumulation (c.f. Frey et al., 2021; Mayr & Freund, 2020)	D	.	I	.
Selective engagement theory (cf. Hess et al., 2021)	.	.	.	I

Table 2 Continued

Theoretical Account	Preference			
	Risk	Time	Social	Effort
Self-efficacy theory (cf. Hess et al., 2021)	.	.	.	I/D
Social-investment theory (cf. Frey et al., 2021)	D	.	.	.
Socio-emotional selectivity theory (cf. Frey et al., 2021; Seaman et al., 2022; Sparrow et al., 2021)	D	I	I	.
Time perception (cf. Seaman et al., 2022)	.	D	.	.

Note. I: Increased risk taking/temporal discounting/altruism/effort discounting with age. D: Decreased risk taking/ temporal discounting /altruism/effort discounting age. U: U-shaped relation between risk taking/temporal discounting/altruism/effort discounting and age. ∩: Inverse U-shape relation between risk taking/temporal discounting/altruism/effort discounting and age. Dot: Not applicable

Supplementary Appendix

Aging and Economic Preferences: Cumulative meta-analyses of age differences in risk, time, social, and effort preferences

Alexandra Bagaïni, M.Sc.[1], Yunrui Liu, M.Sc.[1], Arzie Bajrami, B.Sc.[1], Gayoung Son, B.Sc.[1], Loreen Tisdall, Ph.D.[1], and Rui Mata, Ph.D.[1]

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

Supplementary Appendix

Aging and Economic Preferences: Cumulative meta-analyses of age differences in risk, time, social, and effort preferences

Supplementary Methods

Scoping Review: Literature Search

We first conducted a computerized literature search of publication records on Web of Science to identify previous meta-analyses of age differences in either risk, time, social, or effort preferences. We searched for publications published until November 1st, 2022 that pertained to the specified search terms (Table S2). From our search for meta-analytical studies, we selected those that reported findings on (a) behavioral tasks involving monetary transactions (real or hypothetical), (b) the adult population (i.e., 18 years or above) and (c) economic preferences that met the definitions from Table 1. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021), and details of this search and screening process are available in Figure S1. In a second step, we examined more closely the search strategy, the eligibility criteria and the included studies of the selected meta-analytical studies to inform our search and screening of primary studies. We noted that the meta-analyses identified via the scoping review were heterogeneous, notably with regards to year of publication and eligibility criteria. Therefore, we adapted our search strategy of individual studies such that the meta-analyses could be comparable across economic preferences.

Publication Bias

To explore evidence of publication bias (i.e., tendency to publish only significant effects), we produced for each set of effect sizes a contour-enhanced funnel plot (Peters et al., 2008). This plot displays the distribution of effect sizes against a precision metric (Figure S13). To assess the absence or presence of publication bias in each set of effect sizes, we visually inspected the funnel plots and conducted multilevel Egger's regression tests (Rodgers & Pustejovsky, 2021). For the Egger's tests, we fitted three-level

meta-regression models with different precision metrics, one with standard error and another with the inverse sample size as predictor. In addition, we conducted p-curve tests (Simonsohn et al., 2014) using the dmetar package (Harrer et al., 2019). This test is used to detect evidence of p-hacking; that is, researchers selectively choosing or analyzing data such that non-significant effects become significant (Figure S14). Only effect sizes significant at alpha = 5% level, which determine whether the distribution of p-values is right-skewed distribution and whether studies are properly powered, are included in these analyses (Simonsohn et al., 2014).

Citations

We investigated how the impact of publications on age differences in economic preferences changed over time, and the association between citations and effect sizes and sample sizes. First, for each publication we obtained the number of citations it received every year since it was published, including its citations as a pre-print. Then, for each preference we fitted two linear models, to assess (a) the effect of a publication's average sample size (log-transformed), and (b) aggregated effect size (accounting for effect size dependency) on the median yearly citations (log-transformed), while controlling for the number of decades it has been available (either as a published article or as a pre-print). For these analyses, we included a total of 120 publications (risk = 54, time = 48, social = 12, effort = 6). We excluded publications (n = 6: risk = 3, time = 2, social = 1) for which yearly citation information was not available from Google Scholar.

Supplementary Results

Publication Bias

Risk. From the visual inspection of the funnel plot (Figure S13) and the results of the multilevel Egger's test, there is no evidence of publication bias using either standard error or the inverse sample size as the precision metric (Table S8) nor was there any evidence of p-hacking from the p-curve test. The right skew analyses were all significant, and the flatness tests were all non-significant (Table S9). Lastly, the power

to detect an effect exceeded 80% textcolorred(power = 94%, 95% CI [90.7%, 96.7%]).

Time. results of the multilevel Egger's test showed no evidence of publication bias. (Table S8). P-curve test results suggested no evidence of p-hacking (Table S9), and studies were overall sufficiently powered to detect an effect textcolorred(power = 84% [73.4%-90.8%]).

Social. The multilevel Egger's test results vary depending on the precision metric used: Using standard error as a precision metric we note no significant asymmetry, whereas an assymetry is detected when using the inverse sample size as a predictor (Table S8). When inspecting the p-curve results, we find no evidence of p-hacking (Table S9) and the power to detect an effect on average exceeds 80% textcolorred(power = 83%, 95% CI [58.9%, 94.5%]).

Effort. The results of the multilevel Egger's test showed no evidence of publication bias. (Table S8). P-curve test results suggested no evidence of p-hacking (Table S9), and studies were overall sufficiently powered to detect an effect (power = 99% [97.2%-99%]).

Citations

Figure S15A shows, for each preference, the number of yearly citations of each publication as a function of the number of years it has been published. Figures S15B and S15C show, for each preference, the relation between median yearly citations and the publication's aggregated effect size and average sample size, respectively. Tables S10 and S11 summarize the results from the linear regressions on the association between median yearly citations with (a) effect sizes and (b) sample sizes, respectively.

Risk. We find no significant effect of effect size on the median number of yearly citations. However, older publications and publications with larger samples are more often cited.

Time. There is no significant effect of effect size or sample size on the median number of yearly citations. However, older publications are more often cited.

Social. We note no significant effect of effect size on the median number of yearly citations, but older publications and publications with smaller samples are more

often cited.

Effort. We find no significant effect of effect size or sample size on the median number of yearly citations.

All in all, the results show no evidence that larger effect sizes have received more attention in the literature in the form of citations, which reduces concerns that studies finding larger age differences had a stronger impact in shaping the aging literature on economic preferences.

References

Depping, M. K., & Freund, A. M. (2011). Normal aging and decision making: The role of motivation. *Human Development*, *54*(6), 349–367. https://doi.org/10.1159/000334396

Frey, R., Richter, D., Schupp, J., Hertwig, R., & Mata, R. (2021). Identifying robust correlates of risk preference: A systematic approach using specification curve analysis. *Journal of Personality and Social Psychology*, *120*(2), 538–557. https://doi.org/10.1037/pspp0000287

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2019). Dmetar: Companion R package for the guide "Doing meta-analysis in R".

Harrison, G. W. (1994). Expected Utility Theory and the Experimentalists. In J. D. Hey (Ed.), *Experimental Economics* (pp. 43–73). Physica-Verlag HD. https://doi.org/10.1007/978-3-642-51179-0_4

Jacobs, P., & Viechtbauer, W. (2017). Estimation of the biserial correlation and its sampling variance for use in meta-analysis. *Research Synthesis Methods*, *8*(2), 161–180. https://doi.org/10.1002/jrsm.1218 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1218

Olschewski, S., Rieskamp, J., & Scheibehenne, B. (2018). Taxing cognitive capacities reduces choice consistency rather than preference: A model-based test. *Journal of experimental psychology. General*, *147*, 462–484. https://doi.org/10.1037/xge0000403

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *British Medical Journal*, *372*, n71. https://doi.org/10.1136/bmj.n71

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias

from other causes of asymmetry. *Journal of Clinical Epidemiology*, *61*(10), 991–996. https://doi.org/10.1016/j.jclinepi.2007.11.010

Pustejovsky, J. E. (2014). Converting from d to r to z when the design uses extreme groups, dichotomization, or experimental control. *Psychological Methods*, *19*(1), 92–112. https://doi.org/10.1037/a0033788

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, *26*(2), 141–160. https://doi.org/10.1037/met0000300

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studiesures, and More Online Studies. *Advances in Methods and Practices in Psychological Science*, *2*(2), 107–114. https://doi.org/10.1177/2515245919838781

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Smith, V. L., & Walker, J. M. (1993). Monetary rewards and decision cost in experimental economics. *Economic Inquiry*, *31*(2), 245. https://doi.org/10.1111/j.1465-7295.1993.tb00881.x

Table S1

Overview of moderators included in the meta-regressions.

Moderator	Description	Rationale
Study Design	Two levels: Extreme group design (e.g., young and older adult groups) vs. continuous (e.g., participants aged 18 to 70)	Accounting for potentially more extreme results in studies using extreme group designs
Effect size metric	Two levels: Pearson's r vs. point-biserial correlation	These metrics have different underlying assumptions (Jacobs & Viechtbauer, 2017; Pustejovsky, 2014)
Incentivization	Two levels: hypothetical vs. incentivized	May influence participant's motivation, and may affect the extent that participants reveal their preferences (Harrison, 1994; Smith & Walker, 1993)
Task format	Two levels: description vs. experience (only applicable to risk preferences)	Confound hypothesis (e.g., Frey et al., 2021; Olschewski et al., 2018)
Domain	Three levels: gain vs. loss vs. mixed (only applicable to risk and effort preference)	Motivational theories (e.g., Depping & Freund, 2011)
Study context	Two levels: online vs. in-person	Accounting for changes in how studies are being conducted and how it allows to collect data from larger and more representative samples (e.g., Sassenberg & Ditrich, 2019)
Gender	Proportion of females in the sample	Past literature suggests gender differences in at least some operationalizations of economic preferences (cf. Frey et al., 2021)
Age range	Differences in decades between the young and old sample	A wider age range may increase power to detect age differences

Table S2

Search terms used to conduct the computerized literature searches on Web of Science.

Section	Search Terms
Meta-Analyses of Aging and Preferences	*(age OR aging OR ageing OR "older adults") AND ("risky choice" OR "risk taking" OR "risk-taking" OR altruis* OR prosoci* OR philanthrop* OR generativity OR framing OR "prospect theory" OR "dictator game" OR "delay aversion" OR "delay of gratification" OR "social preference" OR "risk aversion" OR "time preference" OR "intertemporal choice" OR "temporal discounting" OR "delay discounting" OR "effort discounting" OR "effort-based decision" OR "effort-based choice") AND ("meta analysis" OR "meta-analysis")*
Risk	*(age OR aging OR ageing OR "older adults") AND (risky OR "risky choice" OR "risk taking" OR "risk-taking" OR framing OR "prospect theory") AND ("decision making")*
Time	*(age OR aging OR ageing OR "older adults") AND ("temporal discounting" OR "intertemporal" OR "delay discounting" OR "inter-temporal" OR "delay aversion" OR "delay of gratification" OR "delay gratification" OR "time preference*") AND ("choice*" OR "task" OR "decision" OR "game" OR "procedure" OR "measure" OR "paradigm")*
Social	*(age OR aging OR ageing OR "older adults") AND ("altruis*" OR "social*" OR "prosocial*") AND ("dictator game" OR "discounting" OR "moral decision" OR "giving game" OR "economic decision")*
Effort	*(age OR aging OR ageing OR "older adults") AND ("effort") AND (discount* OR decision OR choice OR "tradeoff" OR "trade off" OR "cost-benefit") AND (task OR exert OR game OR paradigm)*

Table S3

Primary study eligibility criteria.

Aspect	**Preference**			
	Risk	Social	Time	Effort
Definition	Decision under risk only, not confounded by a social, time or effort dimension.	Altruism/prosocial decisions only, not confounded by a time, risk or effort dimension.	Decision involving a delay, not confounded by a social, risk or effort dimension.	Decision involving effort, not confounded by a social, temporal or risk dimension.
Domain	Gain, loss and mixed domain	Not applicable	Gain domain only	Gain, loss and mixed domain
Type of measure	Studies with a behavioral measure involving money/rewards (real or hypothetical)			
Decision environment	Studies completed in a laboratory or online or controlled setting. We exclude behavior collected in an MRI scanner, during EEG measurements or in the context of a pharmacological study.			
Population	Healthy adults			
Age	Adults (i.e., majority of participants are at least 18 years old). Sample needs to have an age range of at least 25 years, (i.e., difference between the maximum and minimum age)			
Type of study	Empirical study. Longitudinal or cross-sectional study			
Type of DV	Numerical or graphical format of results. Quantitative value of age differences with data either for each age group or on the relation between behavior and age (e.g., correlation). Excludes categorical outcomes. Data collected under conditions that should be free of experimental manipulations that would result in a confound (e.g., participants shown a prime prior to making decisions)			

Table S4

Three-level meta-regression results (rho = .5)[1] with effect sizes of primary studies on age differences in risk preference (k = 193).

Reg.#	Moderator	Estimate	SE	t-val	p-val	95% CI
1	Age range	<0.001	0.014	0.035	0.972	[-0.029, 0.03]
2	Prop. female	-0.03	0.235	-0.126	0.902	[-0.562, 0.503]
3	Metric-Correlation	-0.016	0.019	-0.864	0.396	[-0.055, 0.023]
3	Metric-Point-Biserial corr.	-0.03	0.033	-0.891	0.379	[-0.097, 0.038]
4	Study-Age continuous	-0.038	0.02	-1.901	0.076	[-0.081, 0.005]
4	Study-Extreme group	-0.018	0.027	-0.67	0.507	[-0.074, 0.037]
5	Incentivization-Hypothetical	-0.024	0.025	-0.983	0.339	[-0.076, 0.028]
5	Incentivization-Incentivized	0.002	0.028	0.088	0.93	[-0.054, 0.058]
6	Domain-Gain	-0.054	0.023	-2.354	0.023	[-0.101, -0.008]
6	Domain-Loss	0.019	0.044	0.42	0.679	[-0.074, 0.111]
6	Domain-Mixed	0.012	0.036	0.345	0.734	[-0.063, 0.088]
7	Task-Description	-0.015	0.024	-0.634	0.529	[-0.064, 0.034]
7	Task-Experience	-0.051	0.039	-1.288	0.217	[-0.134, 0.033]
8	Context-In-person	-0.03	0.026	-1.157	0.254	[-0.082, 0.022]
8	Context-Online	-0.004	0.023	-0.178	0.862	[-0.054, 0.046]

Reporting cluster-robust standard errors and confidence intervals.

[1] Results were not substantially affected by varying rho values (i.e., correlation between sampling errors within studies).

Table S5

Three-level meta-regression results (rho = .5)[1] with effect sizes of primary studies on age differences in time preference (k = 125).

Reg.#	Moderator	Estimate	SE	t-val	p-val	95% CI
1	Age range	0.015	0.01	1.452	0.16	[-0.006, 0.035]
2	Prop. female	0.035	0.150	0.230	0.822	[-0.287, 0.356]
3	Metric-Correlation	-0.042	0.016	-2.694	0.009	[-0.074, -0.011]
4	Metric-Point-Biserial corr	-0.025	0.056	-0.446	0.662	[-0.145, 0.095]
5	Study-Age continuous	-0.045	0.016	-2.878	0.006	[-0.077, -0.014]
5	Study-Extreme group_group	-0.015	0.054	-0.274	0.788	[-0.129, 0.099]
6	Incentivization-Hypothetical	-0.034	0.016	-2.148	0.035	[-0.065, -0.002]
6	Incentivization-Incentivized	-0.087	0.089	-0.973	0.362	[-0.297, 0.123]
7	Context-In-person	-0.014	0.038	-0.386	0.702	[-0.091, 0.062]
7	Context-Online	-0.052	0.016	-3.307	0.002	[-0.083, -0.02]

Reporting cluster-robust standard errors and confidence intervals.

[1] Results were not substantially affected by varying rho values (i.e., correlation between sampling errors within studies).

Table S6

Three-level meta-regression results (rho = .5)[1] with effect sizes of primary studies on age differences in social preference (k = 28).

Reg.#	Moderator	Estimate	SE	t-val	p-val	95% CI
1	Age range	-0.059	0.046	-1.274	0.241	[-0.166, 0.049]
2	Prop. female	0.031	0.257	0.12	0.913	[-0.867, 0.929]
3	Metric-Correlation	0.054	0.075	0.724	0.491	[-0.121, 0.23]
3	Metric-Point-Biserial corr.	0.17	0.05	3.364	0.011	[0.052, 0.287]
4	Study-Age continuous	0.123	0.05	2.476	0.043	[0.005, 0.242]
4	Study-Extreme group	0.103	0.078	1.319	0.221	[-0.075, 0.28]
5	Incentivization-Hypothetical	0.113	0.082	1.372	0.205	[-0.074, 0.3]
5	Incentivization-Incetivized	0.109	0.048	2.268	0.059	[-0.006, 0.223]
5	Context-In-person	0.121	0.056	2.134	0.053	[-0.002, 0.243]
5	Context-Online	0.085	0.097	0.877	0.447	[-0.227, 0.396]

Reporting cluster-robust standard errors and confidence intervals.

[1] Results were not substantially affected by varying rho values (i.e., correlation between sampling errors within studies).

Table S7

Three-level meta-regression results (rho = .5)[1] with effect sizes of primary studies on age differences in effort discounting (k = 23).

Reg.#	Moderator	Estimate	SE	t-val	p-val	95% CI
1	Age range	-0.121	0.217	-0.56	0.616	[-0.825, 0.582]
2	Prop. female	-1.957	2.709	-0.723	0.542	[-13.114, 9.2]
3	Metric-Correlation	0.054	0.193	0.281	0.805	[-0.776, 0.884]
3	Metric-Point-Biserial corr.	0.36	0.128	2.8	0.069	[-0.052, 0.771]
4	Effort Type-Cognitive	0.469	0.023	20.35	0.001	[0.385, 0.552]
4	Effort Type-Physical	-0.086	0.056	-1.533	0.272	[-0.342, 0.17]
5	Domain-Gain	0.252	0.112	2.258	0.065	[-0.021, 0.526]
5	Domain-Loss	0.061	0.106	0.581	0.584	[-0.201, 0.324]

Reporting cluster-robust standard errors and confidence intervals.

[1] Results were not substantially affected by varying rho values (i.e., correlation between sampling errors within studies).

Table S8

Egger's regression test results with effect sizes of primary studies on age differences in risk (k = 193), time (k = 125), social (k = 28) and effort (k = 23) preference. Three-level meta-regression with standard error or inverse sample size as a predictor.

Precision	Estimate	SE	t-val	p-val	95% CI
			Risk		
Standard error	-0.033	0.4	-0.083	0.934	[-0.859, 0.792]
Inverse Sample Size	1.035	3.836	0.27	0.79	[-6.942, 9.012]
			Time		
Standard error	-0.697	0.374	-1.861	0.071	[-1.456, 0.063]
Inverse Sample Size	-4.133	2.536	-1.63	0.115	[-9.338, 1.072]
			Social		
Standard error	3.033	1.987	1.527	0.163	[-1.5, 7.566]
Inverse Sample Size	16.078	6.758	2.379	0.043	[0.656, 31.499]
			Effort		
Standard error	-1.323	2.434	-0.543	0.638	[-11.049, 8.404]
Inverse Sample Size	16.249	13.346	1.218	0.315	[-27.641, 60.14]

Reporting cluster-robust standard errors and confidence intervals.

Table S9

P-curve analysis results. (R)ight-(S)kewness and flatness test for effect sizes of primary studies on age differences in risk (k = 193), time (k = 125), social (k = 28) and effort (k = 23) preference.

Test name	pBinomial	zFull	pFull	zHalf	pHalf	kFull	kHalf
			Risk				
Right-S test	< 0.001	-17.64	< 0.001	-17.465	< 0.001	65(33.7%)	55(28.5%)
Flatness test	0.996	11.417	> 0.999	17.19	> 0.999	65(33.7%)	55(28.5%)
			Time				
Right-S test	0.001	-10.348	< 0.001	-11.354	< 0.001	40(32%)	30(24%)
Flatness test	0.746	5.728	> 0.999	11.222	> 0.999	40(32%)	30(24%)
			Social				
Right-S test	0.212	-5.166	< 0.001	-5.135	< 0.001	14(50%)	9(32.1%)
Flatness test	0.37	2.899	0.998	6.721	> 0.999	14(50%)	9(32.1%)
			Effort				
Right-S test	< 0.001	-11.293	< 0.001	-10.422	< 0.001	11(47.8%)	11(47.8%)
Flatness test	> 0.999	8.061	> 0.999	9.152	> 0.999	11(47.8%)	11(47.8%)

Table S10

Linear regression analysis results for the association between median yearly citations (log scale) and aggregated effect sizes, controlling for the number of years a publication has been cited. Separate results for risk (publications = 54), time (publications = 48), social (publications = 12) and effort (publications = 6) preference.

Predictor	Estimate	SE	t-val	p-val	95% CI
			Risk		
Intercept	0.5	0.046	10.759	<0.001	[0.409, 0.592]
Effect size	-0.074	0.153	-0.484	0.629	[-0.376, 0.227]
Decades in-print	0.142	0.025	5.632	<0.001	[0.093, 0.192]
			Time		
Intercept	0.442	0.037	11.929	<0.001	[0.369, 0.515]
Effect size	0.067	0.086	0.775	0.439	[-0.102, 0.235]
Decades in-print	0.293	0.021	14.232	<0.001	[0.253, 0.334]
			Social		
Intercept	0.261	0.098	2.662	0.01	[0.065, 0.457]
Effect size	0.252	0.248	1.016	0.314	[-0.245, 0.749]
Decades in-print	0.564	0.102	5.506	<0.001	[0.359, 0.769]
			Effort		
Intercept	0.518	0.121	4.283	<0.001	[0.268, 0.768]
Effect size	0.303	0.241	1.259	0.221	[-0.195, 0.801]
Decades in-print	0.636	0.136	4.69	<0.001	[0.356, 0.917]

Table S11

Linear regression analysis results for the association between a publication's median yearly citation count (log scale) and its average sample size, controlling for the number of years the publication has been cited. Separate results for risk (publications = 54), time (publications = 48), social (publications = 12) and effort (publications = 6) preference.

Predictor	Estimate	SE	t-val	p-val	95% CI
			Risk		
Intercept	0.026	0.123	0.215	0.83	[-0.215, 0.268]
Sample size (log)	0.195	0.047	4.159	<0.001	[0.103, 0.287]
Decades in-print	0.183	0.026	7.009	<0.001	[0.132, 0.234]
			Time		
Intercept	0.368	0.079	4.645	<0.001	[0.212, 0.524]
Sample size (log)	0.031	0.027	1.132	0.258	[-0.023, 0.084]
Decades in-print	0.293	0.02	14.837	<0.001	[0.255, 0.332]
			Social		
Intercept	0.845	0.284	2.978	0.004	[0.277, 1.413]
Sample size (log)	-0.253	0.121	-2.095	0.041	[-0.494, -0.011]
Decades in-print	0.508	0.102	4.996	<0.001	[0.304, 0.711]
			Effort		
Intercept	0.694	0.761	0.912	0.371	[-0.88, 2.269]
Sample size (log)	-0.085	0.38	-0.223	0.826	[-0.87, 0.701]
Decades in-print	0.693	0.148	4.679	<0.001	[0.387, 1]

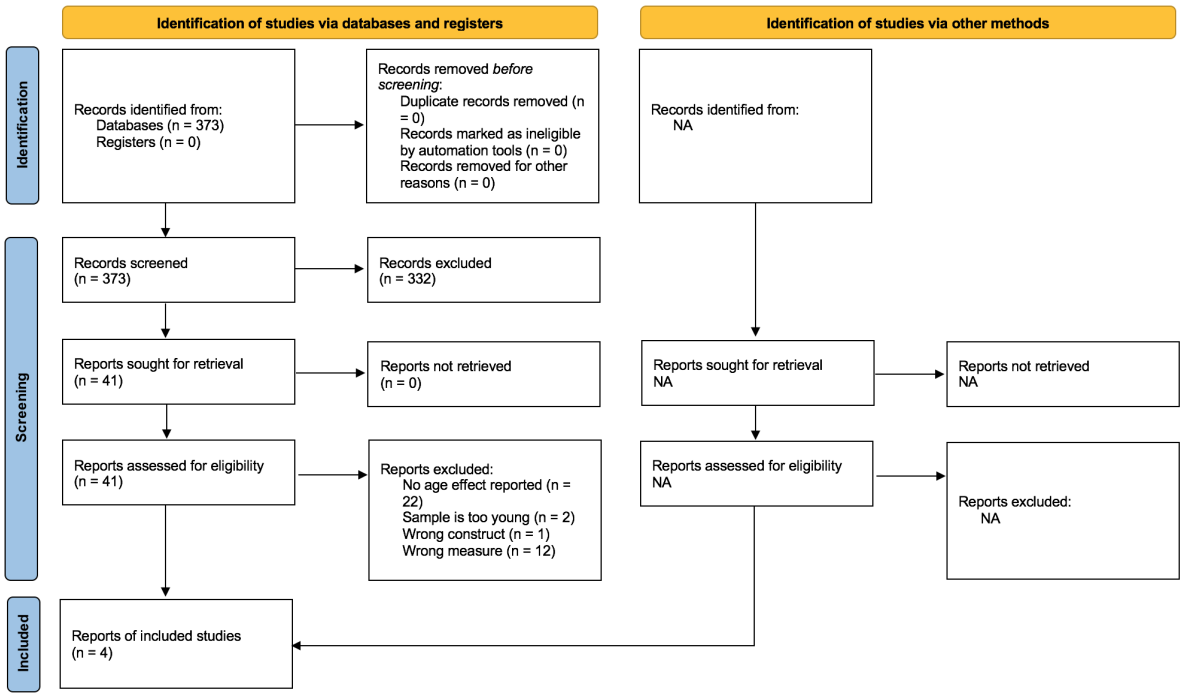


Figure S1

PRISMA flow diagram of the selection process of research synthesis on economic preferences and aging.

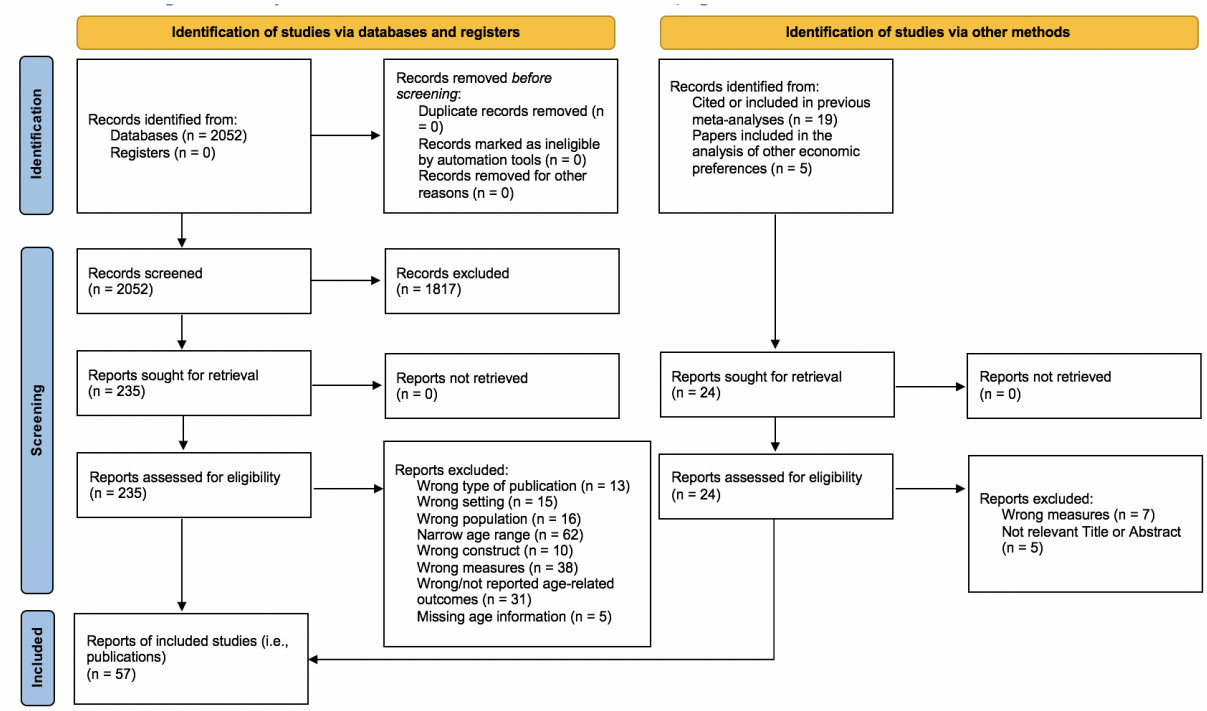


Figure S2

PRISMA flow diagram of the selection process of individual studies on the association between risk preference and age.

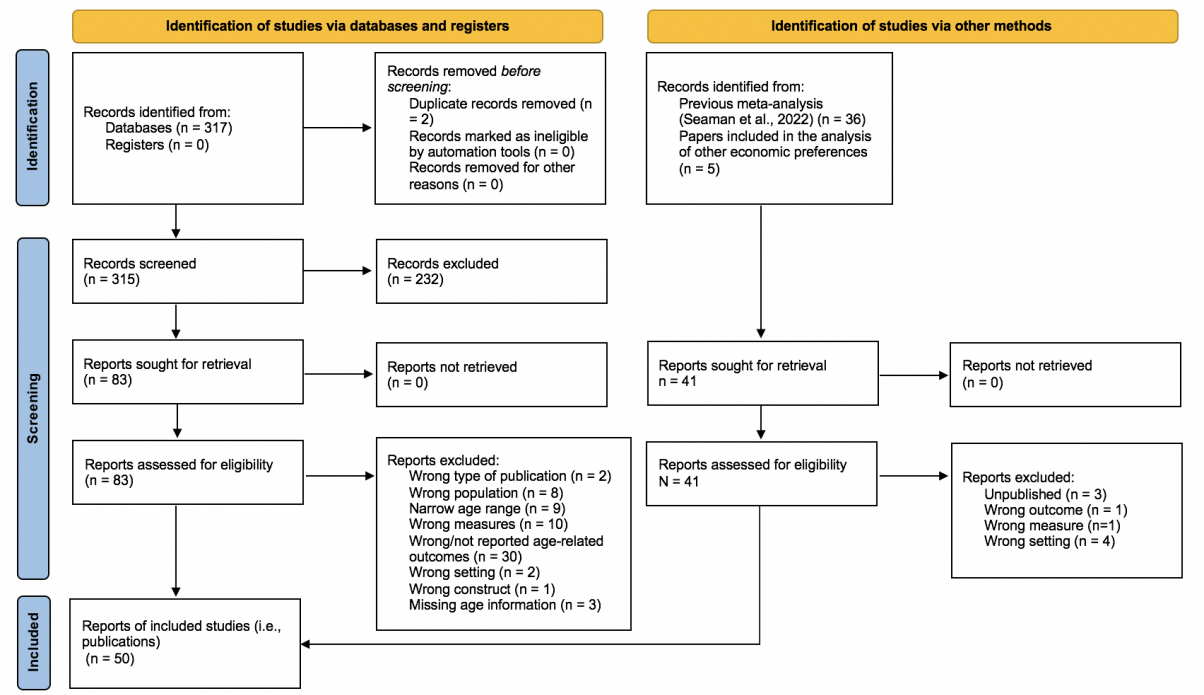


Figure S3

PRISMA flow diagram of the selection process of individual studies on the association between time preference and age.

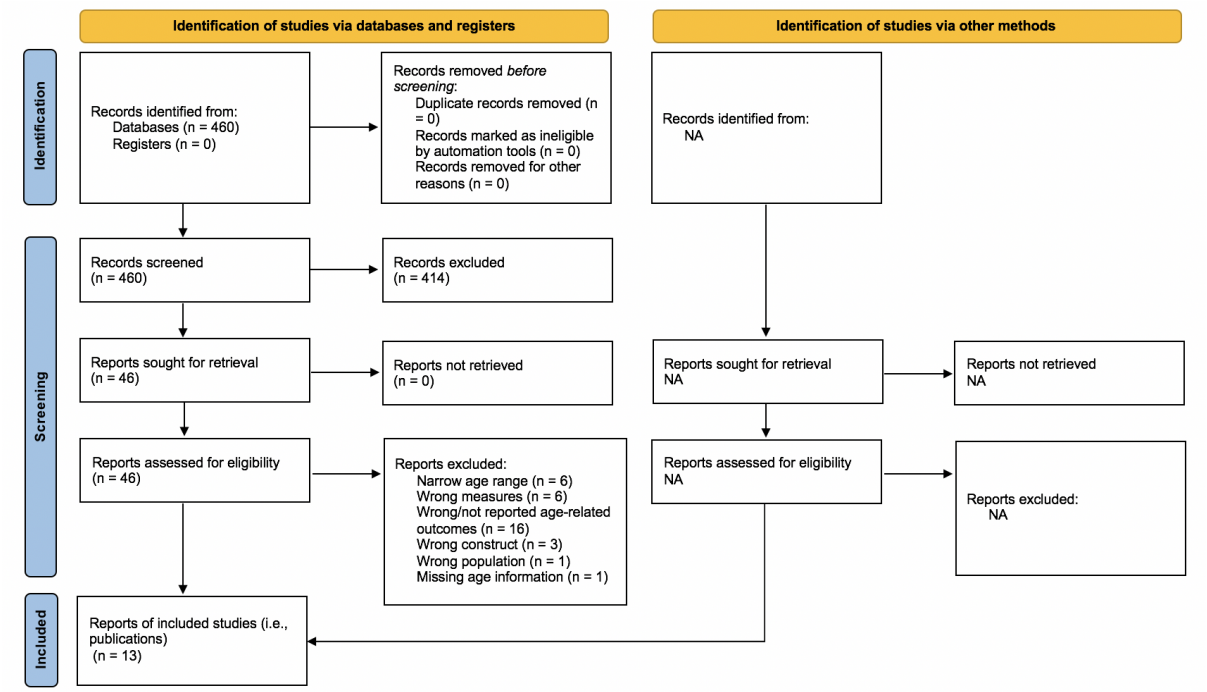


Figure S4

PRISMA flow diagram of the selection process of individual studies on the association between social preference and age.

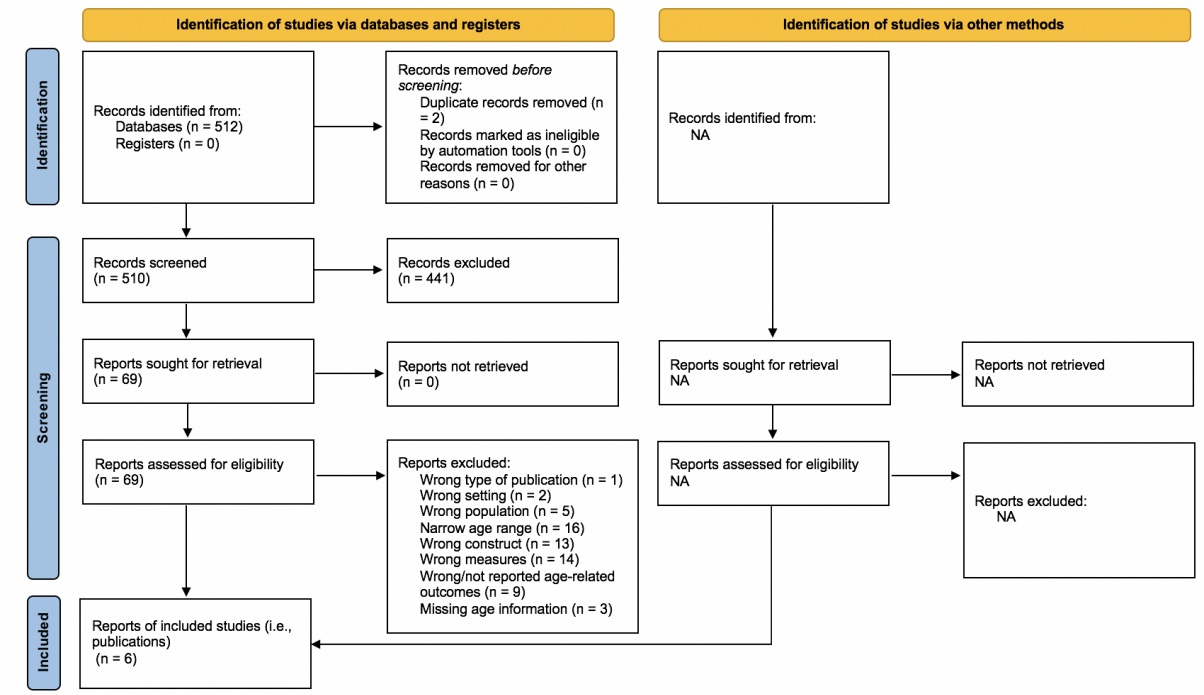


Figure S5

PRISMA flow diagram of the selection process of individual studies on the association between effort-related preference and age.

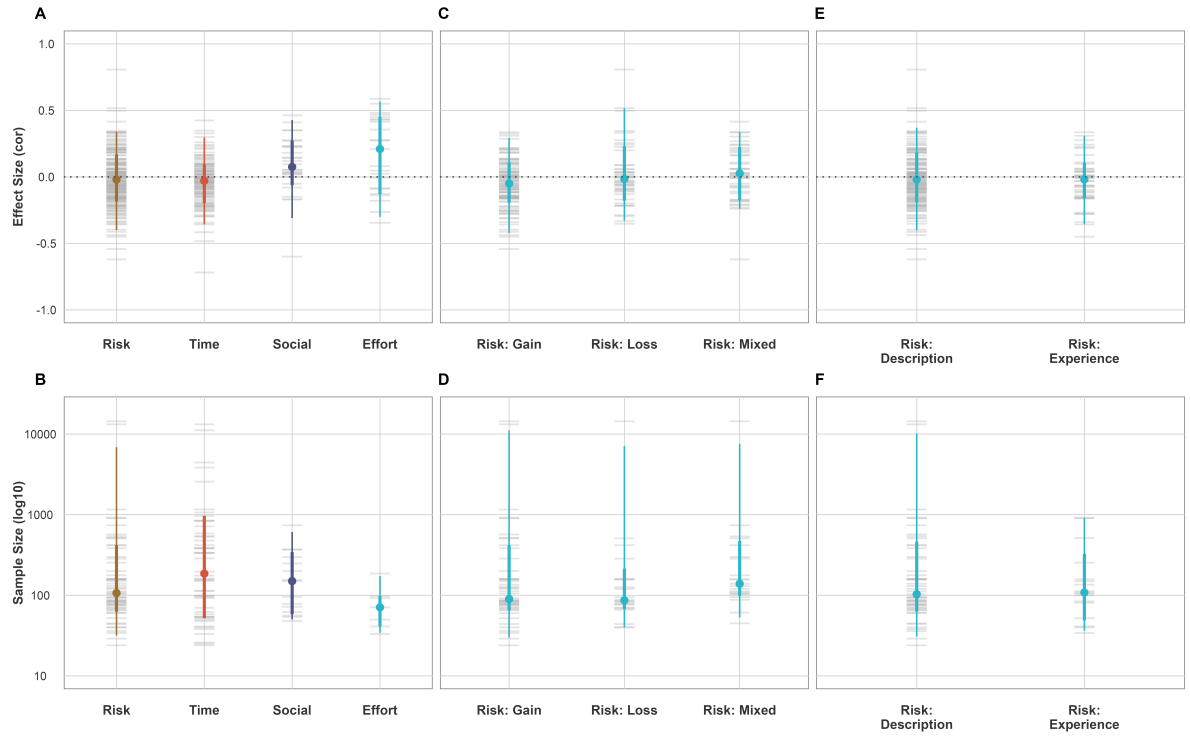


Figure S6

Overview of effect sizes and sample sizes. Grey dashes represent an individual effect size or study, with the overall median, and the 66% and 95% CI. A) Distribution of individual age effects by preference (risk (k = 193), time (k = 125), social (k = 28), and effort (k = 23)). B) Distribution of study sample sizes by preference (risk (studies = 62), time (studies = 54), social (studies = 15), and effort (studies = 7)). C) Distribution of individual age effects by risk preference domain (gain (k = 106), loss (k = 46) and mixed (k = 41)). D) Distribution of study sample sizes by risk preference domain (gain (studies = 48), loss (studies = 22) and mixed (k = 21)). E) Distribution of age effects by risk-taking task category (description (k = 147) and experience (k = 44)). F) Distribution of study sample sizes by risk-taking task category (description (studies = 51) and experience (studies = 17))

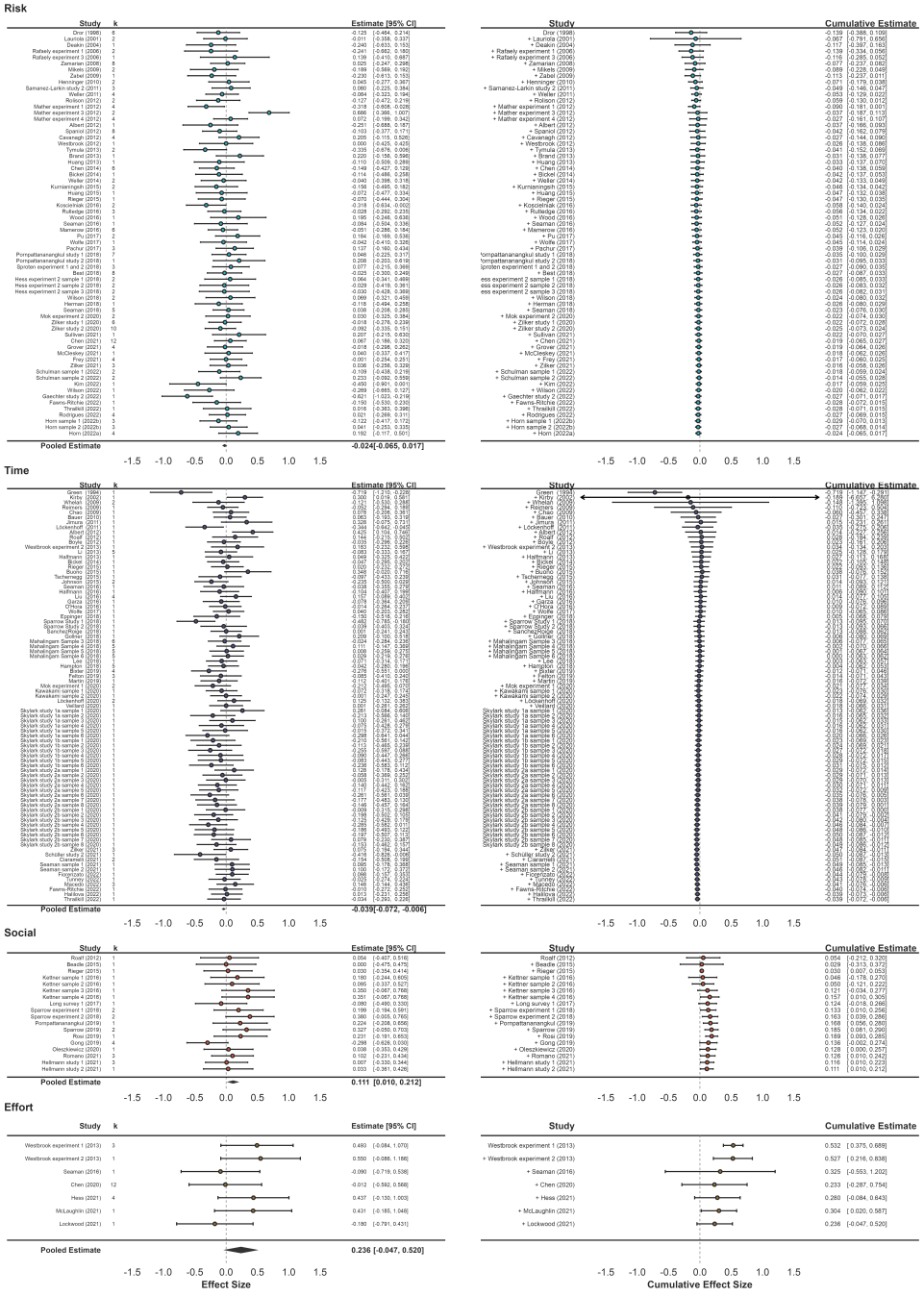


Figure S7

Forest plots of effect sizes on the association between age and risk (k = 193), time (k = 125), social (k = 28), and effort (k = 23) preference. Aggregated forest plots of the three-level meta-analytic model with effect sizes pooled by study with 95% CI, ordered by publication year. Cumulative effect sizes and 95% CI by adding studies in order of publication year.

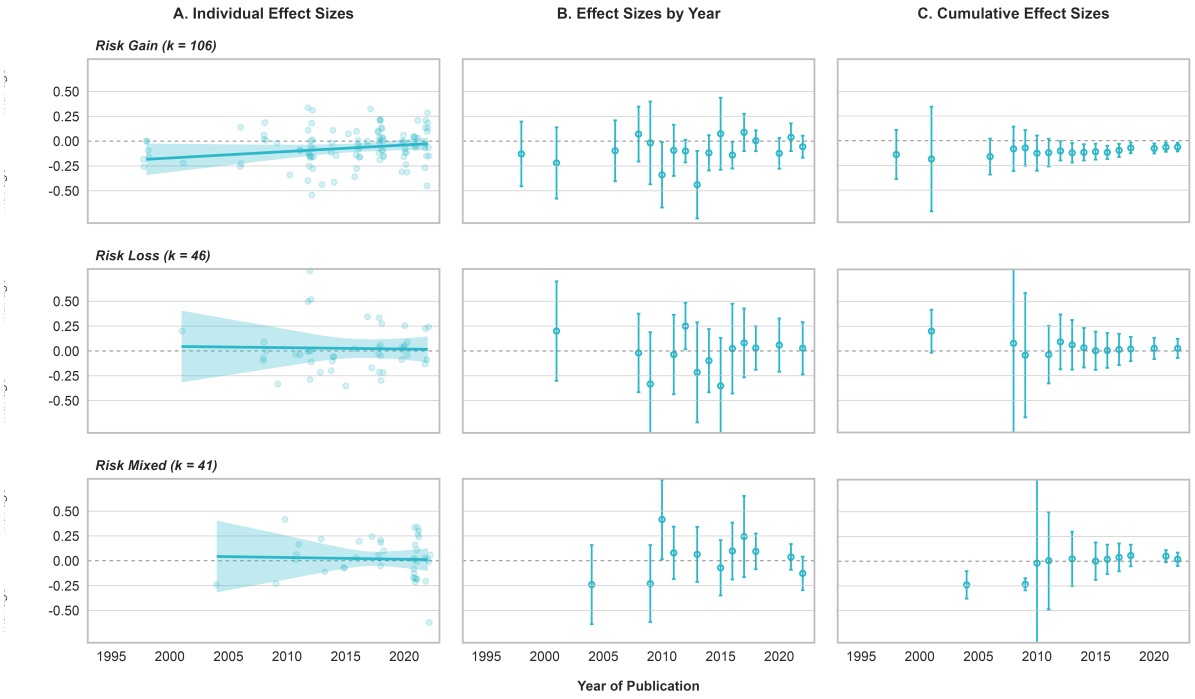


Figure S8

Meta-analytic results of the effect of age on risk taking in the gain (k = 106), loss (k = 46) and mixed (k = 41) domain. A) Scatter plots of the individual effect sizes plotted as a function of the publication year with model predictions and 95% CI. B) Forest plots of the three-level meta-analytic model with effect sizes pooled by year with 95% CI. C) Forest plots of the cumulative effect sizes and 95% CI by year of publication.

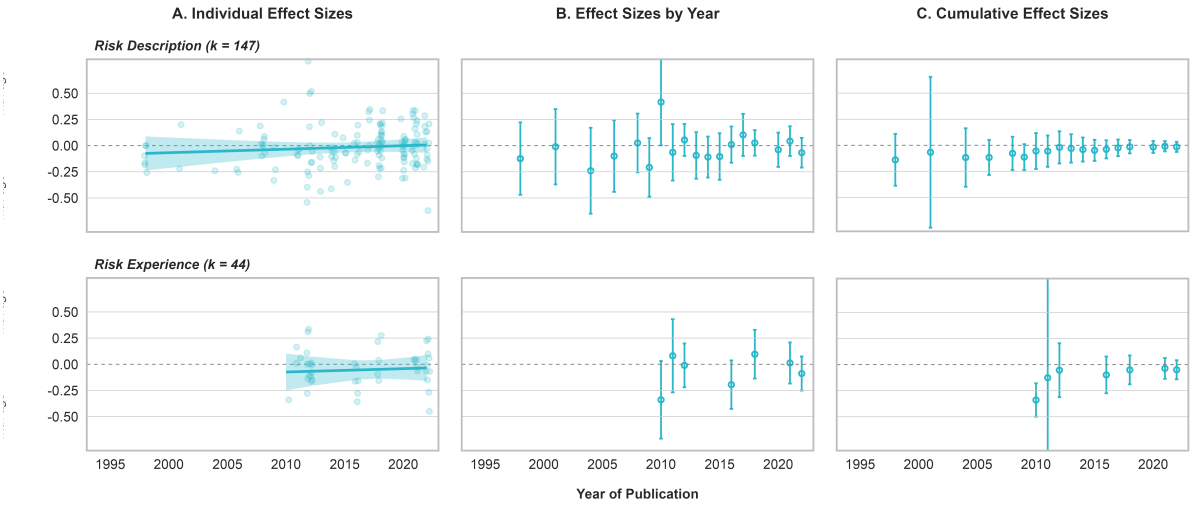


Figure S9

Meta-analytic results of the effect of age on risk taking in decisions from description (k = 147) or experience (k = 44). A) Scatter plots of the individual effect sizes plotted as a function of the publication year with model predictions and 95% CI. B) Forest plots of the three-level meta-analytic model with effect sizes pooled by year with 95% CI. C) Forest plots of the cumulative effect sizes and 95% CI.

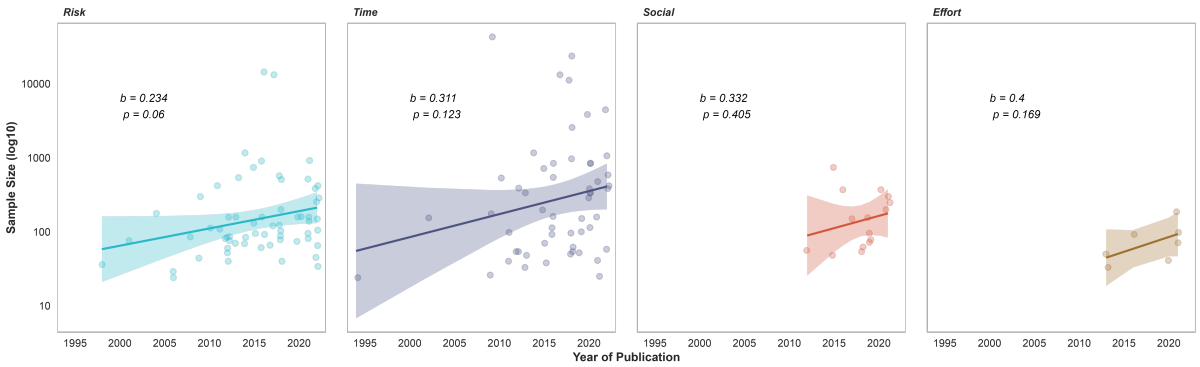


Figure S10

Association between the number of decades a study has been published as of 2022 (transformed into year of publication for plotting purposes) and study sample size for studies on age differences in risk (studies = 62), time (studies = 54), social (studies = 15), and effort (studies = 7) preferences. With model predictions and 95% CI. The beta value and the p-value are results of the linear regression with number of decades since publishing (year of publication - 2022) as a predictor.

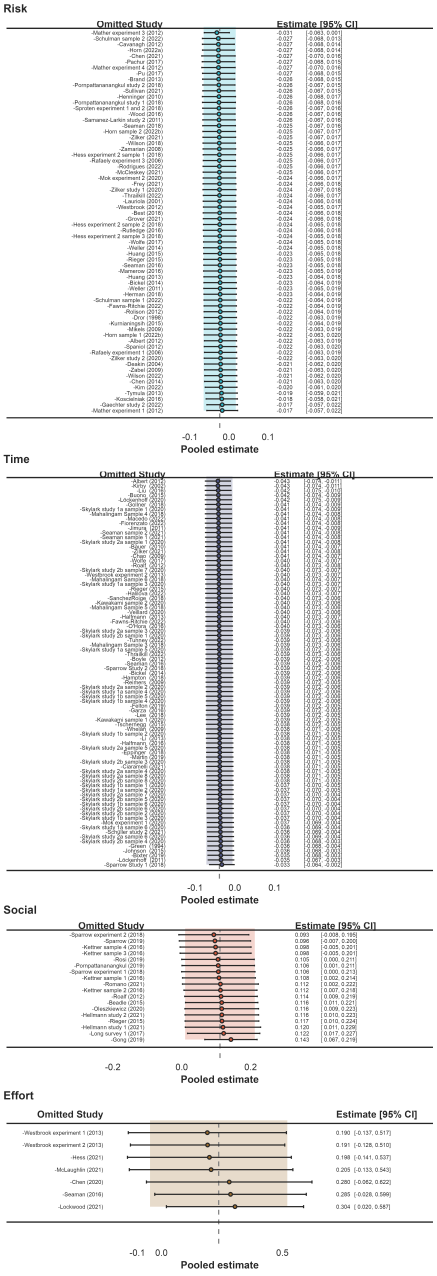


Figure S11

Leave-one-out sensitivity analysis for the three-level meta-analytic model on age differences in risk (k = 193), time (k = 125), social (k = 28), and effort (k = 23) preferences calculated by study. The highlighted section indicates the 95% CI, and the dotted line the mean of the pooled estimate from the three-level meta-analytic model.

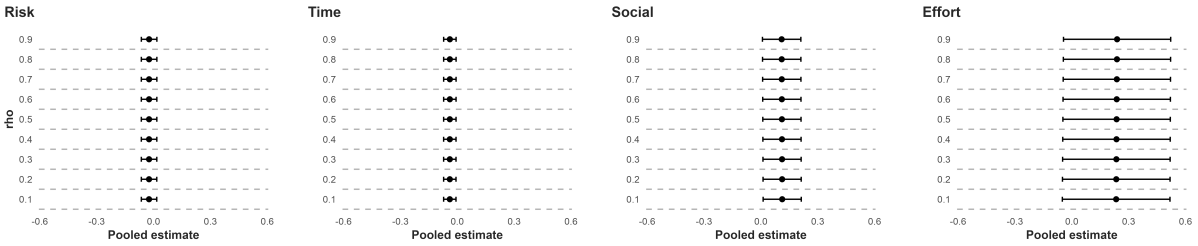


Figure S12

Plots of the pooled estimate for risk (k = 193), time (k = 125), social (k = 28), and effort (k = 23) preferences from the three-level meta-analytic model for different values of rho (i.e., correlation of sampling errors within studies).

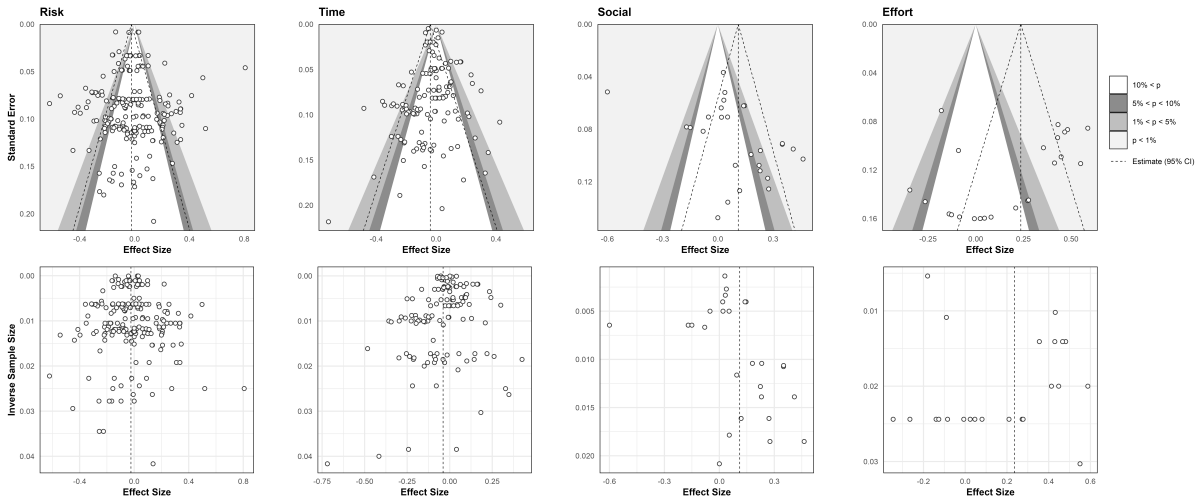


Figure S13

Funnel plots and contour-enhanced funnel plots of the effect sizes of primary studies on age differences in risk (k = 193), time (k = 125), social (k = 28), and effort (k = 23) preferences versus their standard error (upper) and inverse sample size (lower). The shaded regions of the contour-enhanced funnel plot indicate areas of statistical significance, and the white region represents non-statistical significance. The vertical line corresponds to the summary effect size estimate from the three-level meta-analytic model.

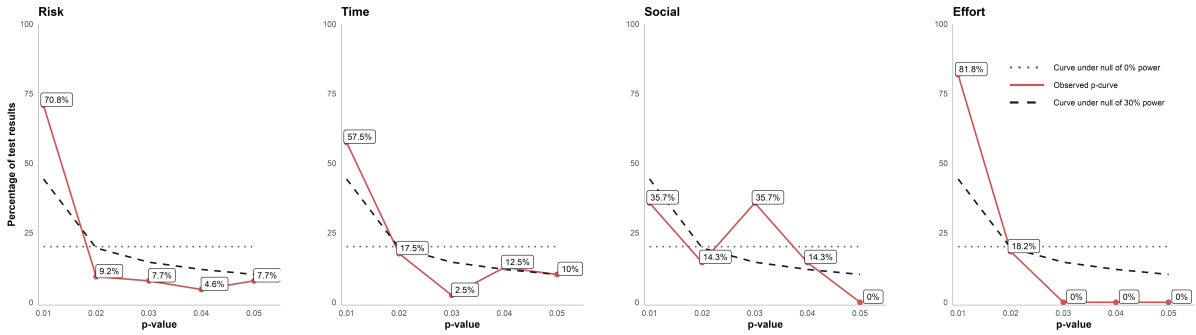


Figure S14

P-curve analysis with effect sizes of primary studies on age differences in risk (k = 193), time (k = 125), social (k = 28), and effort (k = 23) preference.

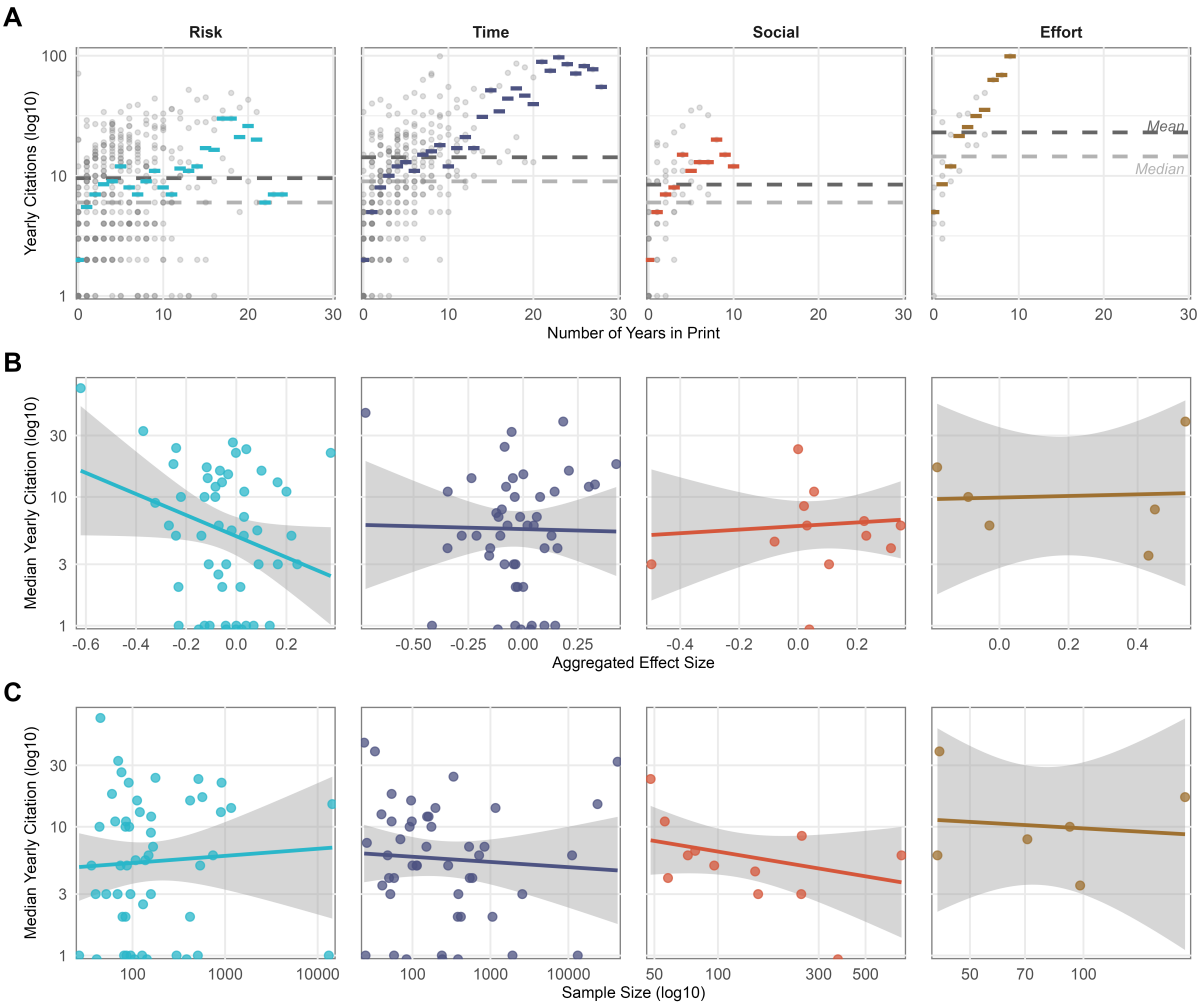


Figure S15

Association between citation count with number of years a publication has been available, the aggregated effect size and average sample size for publications on age differences in risk (publications = 54), time (publications = 48), social (publications = 12), and effort (publications = 6) preference. A) Yearly number of citations as a function of the number of years a publication has been in print. Thick colored dashes represent the median number of citations across all publications (grey dots) for each year. Dark and light grey dashed lines show the overall mean and median number of yearly citations, respectively. B) Scatter plots of the median yearly citation for each publication and its aggregate effect size, with a best fit line and 95%CI. C) Scatter plots of the median yearly citation for each publication and its mean sample size, with a best fit line and 95%CI.

Appendix C

Manuscript 3

Liu, Y., Bagaïni, A., Son, G., Kapoor, M., & Mata, R. (2023). Life-course trajectories of risk-taking propensity: A coordinated analysis of longitudinal studies. *The Journals of Gerontology: Series B, 78*(3), 445-455. doi:10.1093/geronb/gbac175

Life-course trajectories of risk-taking propensity: A coordinated analysis of longitudinal studies

Yunrui Liu, M.Sc.[1], Alexandra Bagaïni, M.Sc.[1], Gayoung Son, B.Sc.[1], Madlaina Kapoor, M.Sc.[1], and Rui Mata, Ph.D.[1]

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

Author Note

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

Abstract

Objectives: How does risk preference change across the life span? We address this question by conducting a coordinated analysis to obtain the first meta-analytic estimates of adult longitudinal age differences in risk-taking propensity in different domains.

Methods: We report results from 26 longitudinal samples (12 panels; 187,733 unique respondents; 19 countries) covering general and domain-specific risk-taking propensity (financial, driving, recreational, occupational, health) across three or more waves.

Results: Results revealed a negative relation between age and both general and domain-specific risk-taking propensity. Furthermore, females consistently reported lower levels of risk taking across the life span than males in all domains but there is little support for the idea of an age by gender interaction. Although we found evidence of systematic and universal age differences, we also detected considerable heterogeneity across domains and samples.

Discussion: Our work suggests a need to understand the nature of heterogeneity of age differences in risk-taking propensity and recommends the use of domain-specific and population estimates for applications interested in modeling heterogeneity in risk preference for economic and policy-making purposes.

keywords: age differences, risk taking, domain specificity, coordinated analysis, life span development

Life-course trajectories of risk-taking propensity: A coordinated analysis of longitudinal studies

People's preferences and attitudes towards risk have the potential to affect many life outcomes, including individuals' labor-market participation, migration, financial investment, and health choices (e.g., Barseghyan et al., 2018; Clark & Lisowski, 2017; Dohmen et al., 2011). As a consequence, understanding individual and age differences in risk preference has been a central concern in psychology and economics for decades (e.g., Mata et al., 2018; Schildberg-Hörisch, 2018). The empirical findings on the link between age and risk preference are, however, patently mixed (e.g., Best & Charness, 2015; König, 2021; Mata et al., 2011) and extant research is characterized by a number of gaps, including a paucity of longitudinal evidence.

In our work, we contribute to a better understanding of the development of risk preference across the life span by providing the first comprehensive coordinated analysis of longitudinal studies of age differences in risk-taking propensity. Understanding individual and age differences in risk preference not only is of conceptual interest but promises to be of applied relevance in assessing the impacts that global population aging will have on individual and societal levels of health and financial well-being.

Age-related Differences in Risk Taking

A recent review listed seven different theories that make predictions about the link between age and risk taking (see Table 1 in Frey et al., 2021). Some of these theories focus on offering a functional explanation for certain patterns of risky behavior across the life span, such as the increased risk taking observed in adolescence (e.g., Defoe et al., 2015), but are mute about the specific mechanisms involved. For example, life-history and risk-sensitivity theories propose that young adulthood is an important phase in which organisms must compete for and accrue resources and, as a consequence, adolescence is associated with increased risk taking geared towards resource acquisition, followed by a reduction once resources have been accumulated (e.g., Del Giudice et al., 2016; Mata et al., 2016; Mishra, 2014). Other theories focus on specific proximal mechanisms that could be associated with reductions in risk taking with increased age.

For example, some have theorized that age-related decline in dopaminergic function leads to decreased exploration and novelty seeking (e.g., Düzel et al., 2010). Other theories focus on motivational mechanisms and suggest that older age is associated with a focus on positive emotions (Carstensen et al., 2000) or loss aversion (Depping & Freund, 2011), leading to a reduced appetite for risk. Finally, yet other theories emphasize the importance of social roles, such as normative life transitions to adult roles (e.g., getting a job, having children), that lead to systematic changes in personality (e.g., conscientiousness) with consequences for risk taking (Bleidorn et al., 2013).

Despite the variety of theoretical stances, a common thread in the aforementioned perspectives is that they suggest an overall reduction in risk taking past young adulthood and across adulthood and aging. Empirical evidence for such a reduction, however, is mixed. Epidemiological data focusing on causes of death or criminality support this idea (Steinberg, 2013). In turn, previous meta-analyses focusing on behavioral paradigms (Best & Charness, 2015; Mata et al., 2011) have revealed rather heterogeneous patterns of age differences in risk taking, with only some measures or task conditions showing the predicted reduction across age. Evidence has accumulated, however, that age differences are more reliably detected in self-report questionnaires capturing individuals' propensity to take risks in a wide set of domains (e.g., König, 2021). For example, a recent study by Frey et al. (2021) directly compared behavioral (i.e., monetary gambles) and self-report (i.e., risk-taking propensity) measures in a large representative sample and found that self-report, but not behavioral measures, were systematically associated with demographic characteristics, such as age. Moreover, self-report measure have higher convergent validity, and thus higher construct validity than behavioral measures (Frey et al., 2017). Indeed, a number of cross-sectional (e.g., Bonem et al., 2015; Mata et al., 2016), as well as a few longitudinal studies (e.g., Banks et al., 2020; Bonsang & Dohmen, 2015; Dohmen et al., 2017; Josef et al., 2016) suggest a reduction in risk taking-propensity with increased age (see König, 2021, for an overview).

All in all, past results suggest that age reductions in risk taking can be

systematically detected at least when captured by self-reported propensity measures. However, only a few studies have examined whether such age-related patterns hold longitudinally and there is no quantitative meta-analysis of the age-related changes in risk-taking propensity. There are two main reasons why such a synthesis is needed. Firstly, a quantitative synthesis of age differences in risk-taking propensity can help clarify the extent to which an individual's appetite for risk changes systematically with age as well as examine important moderators that have not been thoroughly considered in past work. Indeed, there is still uncertainty concerning the extent to which age patterns differ across populations and geographic regions (e.g., Mata et al., 2016) and are moderated by gender and domain (Falk et al., 2018; Josef et al., 2016; König, 2021). A number of theories have been proposed that imply gender differences in risk taking, with males engaging more in risk-taking activities relative to females (see Frey et al., 2021, for an overview). Two meta-analyses are compatible with this view (Byrnes et al., 1999; Cross et al., 2011) and large-scale studies find pan-cultural evidence for such gender differences (Falk et al., 2018; Falk & Hermle, 2018; Mata et al., 2016). However, the extent to which such gender differences interact with age is less clear (e.g., Josef et al., 2016). Indeed, age by gender interactions could help reveal the extent to which gender-specific mechanisms, be they biological or environmental, play a role in age-related reductions in risk taking across adulthood. A second open issue concerns the role of domain. One qualitative review of domain-specific differences in the patterns of age differences suggests that some domains see more pronounced age effects relative to others, for example, systematic age differences are more pronounced in the physical domain compared with interpersonal domain (König, 2021). However, so far, the magnitude of domain effects has not been assessed quantitatively in a systematic manner, making it difficult to assess to what extent these differences are reliable and merit further theorizing.

Secondly, from an applied perspective, quantitative and robust estimates are important to assess the role of global population aging in individual and societal levels of risk taking in real-world settings, such as financial markets (Barseghyan et al., 2018)

or sustainable consumption (e.g., McCollum et al., 2017). For example, recent modeling efforts of consumer decision-making attempt to integrate fine-grained estimates of individual and group differences in risk preferences to inform expectations about economic growth in the next decades (e.g., McCollum et al., 2018). A quantitative assessment of age-related differences and understanding of their generality across domains, populations, and periods or cohorts will be crucial in developing the next generation of such integrated assessment models, which aim to include population heterogeneity and have become central to policy making (e.g., Trutnevyte et al., 2019).

Overview of The Present Study

As noted above, aging research has identified mixed results concerning the link between age and risk taking, as well as the role of gender and domain-specificity in such age-related patterns. In this study, we aim to use a coordinated and integrative data analysis method to clarify these issues by answering the following specific research questions: 1) What are the overall age patterns of mean-level change in self-reported risk-taking propensity across various data sets? 2) Are there substantive gender differences in these mean age trajectories? and 3) To what extent do age and gender differences vary significantly by domain, such as general and specific domains? Altogether, we contribute to describing age differences in risk preference across the life span by providing the first quantitative summary of age differences in self-reported risk-taking propensity for a comprehensive set of longitudinal panels covering the widest possible set of geographic regions.

For this purpose, we conducted a broad search for longitudinal panels containing self-report measures of risk-taking propensity spanning three or more waves from any publicly available source around the world. We then used a coordinated analysis approach, analyzing independent samples in a harmonized statistical model that optimizes the comparison of results (Hofer & Piccinin, 2009, 2010; Piccinin & Hofer, 2008; Weston et al., 2020). This approach increases comparability and generalizability of results across distinct samples by using the same set of analytic choices and models without, however, assuming equivalence between measures (cf. Graham et al., 2020;

Graham et al., 2022). Specifically, in our study, we first used a multilevel model to capture the association between age and risk-taking propensity across each longitudinal sample and domain. In a second step, we used a meta-analytic approach to integrate the estimates obtained from each sample into summary estimates per domain. This approach allowed us to provide the first quantitative meta-analytic comparisons of the age trajectories of risk-taking propensity across samples and domains. All in all, our approach will contribute to reliable and effective cumulative science in the domain of adult development and aging.

Methods

Data

We identified the largest possible number of longitudinal panels containing at least three waves of self-reported risk-taking propensity, in either general (e.g., How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?) or specific domains (e.g., How do you evaluate your attitude towards risk regarding financial investments/driving a car/leisure time and sport/your occupation/your health/your faith in trusting other people?). General propensity items typically refer to risk taking without any specification of situation or behaviors whereas specific domains indicate specific life matters or target activities. We identified panels and relevant studies by consulting previous literature on the temporal stability and life-span trajectory of risk taking (e.g., König, 2021; Mata et al., 2018; Schildberg-Hörisch, 2018), as well survey listings and data repositories (e.g., Gateway to Global Aging Data, the Integrative Analysis of Longitudinal Studies of Aging and Dementia (IALSA) project).

After identifying potential panels and studies, we systematically checked each one to ensure that it met the following inclusion criteria: 1) included longitudinal data with three or more waves of general or domain-specific risk-taking propensity that were available by March 31st, 2022; 2) included information on the age and gender of the respondents; 3) included an adult population with age range spanning 30 years or

more[1]. To identify risk-taking propensity items, we devised a list of terms related to risk in general (e.g., risk) and specific domains (e.g., driving, recreational activities, health-related behaviors) and searched for these terms in the available variable search engines, codebooks, or questionnaires of each panel. According to a broad schema, we then classified each item as either a general risk-taking propensity measure or a specific measure to one of the following domains: financial, driving, recreational, occupational, health, and social. Details of all the items for each panel and domain are available in our companion website and Github repository.

Panels or studies that included data from several countries (i.e., Preference Parameters Study of Osaka University (GCOE), Survey of Health, Ageing and Retirement in Europe (SHARE)) were treated as separate samples at the country level to avoid confounding potential cross-country differences. Altogether, we identified 12 panels (26 samples) that met our inclusion criteria; specifically, the DNB Household Survey (DHS), the Preference Parameters Study of Osaka University (GCOE), the German Longitudinal Election Study (GLES), the Household, Income and Labour Dynamics in Australia (HILDA), the Health and Retirement Study (HRS), the Life in Kyrgyzstan Study (LIKS), the Panel of Household Finances (PHF), the Sparen und Altersvorsorge in Deutschland (SAVE), the Survey of Health, the Ageing and Retirement in Europe (SHARE), the German Socio-Economic Panel (SOEP), the Understanding America Study (UAS) and the UK Household Longitudinal Survey (USoc). The detailed identification and screening longitudinal panels process can be seen in Figure 1. Table 1 provides an overview of all samples. We also offer a detailed description of each panel in the supplementary materials.

Data Preprocessing

Panels differed in the measures of risk-taking propensity and scales used therefore we performed a series of operations to increase the comparability of risk-taking propensity responses across samples. First, we recorded items such that

[1] For this reason, the National Longitudinal Study of Adolescent to Adult Health (Add Health) and National Longitudinal Survey of Youth 1979 Child and Young Adult (NLSCYA) were excluded from our analyses.

higher scores indicated a higher degree of risk-taking propensity across all measures. Second, some measures relied on an 11-point scale whereas others relied on 4- or 7-point scales, so we transformed all propensity scores using $POMP = (observed - min)/(max - min) * 10$ based on the Percentages of the Maximum Possible (POMP) score method (Cohen et al., 1999). To increase comparability between scales, we also z-transformed scores based on a reference age group (50–55 years old) in each sample. For demographic variables, we centered the respondent's age to a reference age (50 years old) and converted it to decades by dividing it by 10. Gender was dummy coded such that in all samples 1 = female and 0 = male. To explore whether age-related changes in risk-taking propensity varied between younger and older cohorts, age at first assessment was also dummy coded as below or over 60 years old (for a similar approach, see Graham et al., 2020).

Data Analysis

We used R (R Core Team, 2020) for all our analyses. We estimated multilevel models with the lme4 package (Bates et al., 2015) and conducted meta-analyses using the metafor package (Viechtbauer, 2010).

Individual Sample Analysis

The relation between age and risk-taking propensity can theoretically take various forms but past empirical work suggests mostly negative linear or quadratic patterns with age (Dohmen et al., 2017; Frey et al., 2021; Josef et al., 2016; Rolison et al., 2014; Schurer, 2015), with some debates concerning possible interactions with gender (Josef et al., 2016). In our analysis, we compared a number of models to describe the relation between age and risk-taking propensity and possible interactions with gender.

First, we fit an unconditional model (i.e., an intercept-only model) to provide a baseline for comparing subsequent models. The unconditional model allows variance decomposition and comparison of the within- to between-subject variability. The intraclass correlation coefficient (ICC) is calculated by dividing between-subject variance by the sum of the between- and within-subjects variance (i.e., $ICC = \tau_{00}/(\sigma^2 + \tau_{00})$). A

low ICC (< 0.2) indicates less interindividual variability whereas a high ICC (> 0.8) indicates less intraindividual variability. A medium ICC (between 0.2 and 0.8) suggests that there is inter- and intraindividual variability (Landis & Koch, 1977).

Subsequent models estimated the relation between age, gender, and risk-taking propensity in different domains. A second model included age as a predictor but did not consider differences across participants (fixed effects model). A third model, in turn, included age as both a fixed and a random slope. Fourth, we added gender into the regression. In this way, we coded the relation between interindividual differences (age) in the change trajectories and the time-invariant characteristic (gender) of the individual to compare whether age is associated with risk-taking propensity in males and females in the same manner. Fifth, an additional model further included an age by gender interaction. Sixth, we fit a quadratic growth model to assess nonlinear change. We did this by squaring age and entering this into the model. Seventh and finally, we added gender into quadratic growth model to assess potential age differences in the quadratic trajectories.

In summary, for a given criterion, we fitted a possible total of seven models: 1) intercept-only model (M1), 2) age fixed effects model (M2), 3) age fixed and random effects model (M3), 4) age fixed and random effects model with gender (M4), 5) age fixed and random effects model with gender, including an age by gender interaction (M5), 6) age quadratic growth model (M6), and 7) age quadratic growth model with gender (M7). An overview of all models is presented in Table S1 and results for all models are provided in the companion website.

Meta-analysis

After obtaining the estimates of interest for each model and domain in every sample, we computed a summary of each estimate for general, financial, driving, recreational, occupational and health risk-taking by conducting a meta-analysis. We did not conduct a meta-analysis for social risk-taking propensity because only one panel (i.e., SOEP) included the survey item in this domain; therefore, we included the relevant results of social on the companion website rather than in the main text. Each

meta-analysis produced an overall effect size (weighted by the sample size) with corresponding standard errors, confidence intervals, and estimates of heterogeneity (I^2, Q) for each estimate (e.g., age, gender). I^2 indicates the ratio of true heterogeneity to total variance in the observed effects across studies, ranging from 0% to 100%. Heterogeneity can be quantified as low, moderate, and high with upper I^2 limit of 25%, 50%, and 75%, respectively (Borenstein et al., 2017; Higgins et al., 2003). Cochran's Q is computed as the weighted sum of the squared deviations of each study's effect size from the overall pooled estimate. A significant Q-value ($p < 0.05$) indicates heterogeneity in the dispersion of effect sizes. In line with the heterogeneity results, we used random effects models to meta-analyze each set of estimates (Borenstein et al., 2010). To explain the heterogeneity between samples for each model (i.e., M1 - M7) and risk-taking domain, we used a meta-regression with a number of moderators (i.e., continent, mean age, scale range and baseline survey year) to identify the effect of these moderators on outcomes effect. Table 1 offers a description of these moderators for each sample. We only conducted the meta-regression with the additional moderators for general and financial risk taking because other domains had only a small (3) number of samples.

Model Comparison

We compared and selected models based on the results of the meta-analysis and analysis of variance (ANOVA), which revealed by the Akaike Information Criterion (AIC, Akaike, 1998) and the Bayesian Information Criterion (BIC, Schwarz, 1978). First, we compared models (i.e., M3–M5) for samples that included three or more waves, to identify which model could best capture the linear growth change in every domain. Then, we compared models (i.e., M3–M7) for samples including four or more waves, to compare the linear growth against quadratic change. Based on meta-analysis results, in all domains (except for driving), there is no significant quadratic age effect, thus we only report the model comparison results for three or more waves in the main text; however, we also report the model comparison results for four or more waves on the companion website.

Cohort effects

To test whether cohort could account for differences in risk-taking propensity trajectories within each sample, we added the dummy coded cohort variable (baseline age under or above 60 years old) as a level 2 predictor to the best fitting model (M4) resulting in two additional models: one model without any interaction effect (i.e., age fixed and random effects model with gender and dummy cohort, M8) and another considering the interaction effect (i.e., age fixed and random effects model with gender and dummy cohort, and including interactions between age, gender, and dummy cohort, M9). We also conducted model comparisons and meta-analyses of these two additional models. The rationale for adding baseline age group as a level 2 predictor is that this provides one way of estimating whether birth cohort in a given start year is associated with the risk trajectories in each sample (cf., Graham et al., 2020).

Variance decomposition

We estimated a simple multilevel model without predictors (i.e., an intercept-only model) but with specific random effects to allow clustering samples and help estimate the role of different variance components to the age and gender effects obtained from the best fitting models (M4). In this way, we can better understand which variable explains most of the variance in the growth curve.

Results

Panel Identification and Selection

As shown in Table 1, we identified a total of 12 eligible panels, consisting of 26 longitudinal samples and over 180,000 unique respondents. Only a small subset of these data has ever been analyzed to study age differences in risk-taking propensity (cf., Banks et al., 2020; Bonsang & Dohmen, 2015; Dohmen et al., 2017; Josef et al., 2016).

Modeling of Age Effects

Our main goal was to assess the association between age and risk-taking propensity across the life span (18–90 years of age). For this purpose, we tested a number of models that estimated the effects of age for each panel and domain

separately in a total of 42 data sets (26 samples, containing 1 to 7 domains per sample; see Table 1). As described in more detail in the methods section, our approach was to test and compare several models, from a simple intercept-only model that ignores potential age effects to others that considered different manners in which age may be related to risk-taking propensity (e.g., linear, quadratic), as well as others considering potential moderation effects (e.g., gender).

The results of an intercept-only model with ICC values (Figure S1 and Table S2) indicated that approximately 42% (M = 0.42, SD = 0.11) of the total variance in risk-taking propensity across measurement occasions can be attributed to between-person variance. The remainder was attributable to within-person change and measurement error. The considerable within- and between-subject variance warranted following mixed-effects models, which aim to capture any potential systematic within-person changes related to age.

More notably, the mixed-effects model with age and gender as predictors but no age by gender interaction (M4) was the best fitting model for the majority of samples in 32 of 42 comparisons (76%). An overview of model performance across domains is provided in Table S3. Visual inspection of model fits and comparison of regression coefficients across models also suggest that the age by gender interaction effects, when significant, were small in magnitude. As a consequence, in what follows, we maximize comparability across samples and domains by reporting the meta-analytic summarized results for M4. We also provide the results for all additional models and respective meta-analytic summaries on the companion website (https://cdsbasel.github.io/ageriskmeta). Similarly, when considering the additional models aimed to capture cohort effects, the model with age, gender, and dummy cohort as predictors but no interaction provided the more parsimonious fit.

The results per data set and domain can be best observed in Figure 2. All in all, across domains, we detect a negative effect of age. One large source of differences across samples, however, is domain, with some domains showing steeper declines across the adult life span. For example, the age slope observed for recreational and occupational

risk-taking propensity appears more pronounced than for other domains. Concerning gender effects, there are some clear differences between males and females, with the former showing on average higher levels of self-reported risk-taking propensity across domains.

Meta-analytic Estimates

We aimed to integrate the results from various populations and compare domains by computing a meta-analytic summary of the results per domain. The comparison of age and gender effects between domains is detailed in Figure 3 and Table S4. Primarily, we observe an overall negative effect of age, ranging from -.08 to -.18 of a standard deviation per decade (panel a). The age effects per decade are about half to a quarter of the size of those detected for gender, which tend to range between -.25 and -.39 (panel b). Second, both the effects of age and gender are domain-specific in the sense that some domains show larger age and/or gender effects relative to others: the recreational (-0.17, 95% CI = [-0.21, -0.12]) and occupational (-0.18, 95% CI = [-0.20, -0.16]) domains show particularly steep declines across the adult life span whereas driving (-0.39, 95% CI = [-0.45, -0.32]) and recreational (-0.37, 95% CI = [-0.44, -0.30]) show the largest gender effects. Finally, the meta-analytic summary also confirms the existence of a cohort difference in the financial (-0.06, 95% CI = [-0.10, -0.02]), occupational (-0.17, 95% CI = [-0.29, -0.05]) and health domains (-0.10, 95% CI = [-0.13, -0.07]), indicating that older cohorts (i.e., above 60 years old) tended to show steeper slopes in these three domains than younger cohorts.

Estimation of Heterogeneity

Our meta-analytic approach can also be helpful to understand the sources of heterogeneity in adult age differences around the world. We addressed the issue of heterogeneity in two ways. First, for each meta-analysis conducted, we estimated the I^2 statistic (i.e., the ratio of sample heterogeneity to total variability) that is often used to quantify heterogeneity (Higgins & Thompson, 2002). The results across meta-analyses showed relatively large I^2 ($> 95\%$) values, suggesting that more than 95% of the observed variance between studies reflects variance in true effect sizes rather than

sampling error (Borenstein et al., 2017). As a consequence, for each meta-analysis conducted, we tested whether including additional predictors (i.e., continent, mean age, scale range, and baseline survey year) moderated the effect sizes. Overall, adding these predictors did not provide additional explanatory value, suggesting that neither these sample nor measure characteristics contribute systematically to the observed effects and other unobserved characteristics are responsible for the differences between samples.

Second, we conducted a variance decomposition for the best fitting model results across all samples using a simple multilevel model in which we clustered results by sample, domain, continent, scale range and baseline survey year. Concerning age-related differences, the results suggested that domain and sample were responsible for approximately 57% and 25% of the variance observed in age-related differences respectively, with baseline survey year capturing about 10%. In turn, for gender differences, domain and sample were responsible for around 38% and 33% of the observed variance respectively, with baseline survey year capturing around 21% of the variance. Altogether, these results emphasize that differences across domains and samples are sizeable and contribute to a large portion of heterogeneity in age and gender differences in risk-taking propensity. In addition, the effects of baseline survey year suggest that period effects also contribute to some of the differences across samples in risk-taking propensity.

Discussion

In this study, we aimed to describe age-related changes in risk-taking propensity by conducting a coordinated analysis of a large set of representative longitudinal panels from around the world. Using multilevel models to assess sample level changes and meta-analyses to quantify the overall trajectory, we were able to collate data from 26 samples stemming from 19 different countries and spanning up to 29 years to document universal and sample-specific age-related trajectories in general and domain-specific risk-taking propensity.

Our work makes two main contributions. Above all, our results provide the first systematic investigation of age and gender differences in risk-taking propensity across a

large set of longitudinal panels and show strong evidence for an age-related reduction in risk-taking propensity. Age differences in risk-taking propensity are accompanied by gender differences, specifically, males consistently reported higher levels of risk taking than females but we find little evidence for an age by gender interaction. Our work thus expands previous narrative reviews (König, 2021) and quantitative syntheses focusing on behavioral paradigms (e.g., Best & Charness, 2015; Mata et al., 2011), which implied a heterogeneous pattern of age differences across studies and measures, and suggests, instead, a rather universal character of age-related reduction in risk-taking propensity across the adult life span. These findings are compatible with several theories of aging and risk taking that propose a decline in risk-taking propensity across adulthood as well as a number of theories that suggest a gender differential in risk-taking propensity (cf. Cross et al., 2011; Frey et al., 2021).

Second, and notwithstanding the overall age-related reduction in risk-taking propensity, we provide meta-analytic evidence for domain-specificity by demonstrating that some domains, such as recreational and occupational domains, show systematically more pronounced declines with age relative to others, such as driving, health, or financial domains. Also, driving and recreational domains show larger gender differences than others. Previous research categorized risk-taking propensity into two clusters (König, 2021), an interpersonal cluster, including recreational, career/occupational, social and ethical risk-taking, and a second cluster including domains that directly threaten mental and physical well-being, such as financial, driving, health, and environmental risk-taking. Interpreting our results in line with this distinction, risk-taking propensity associated with interpersonal domains showed a steeper age-related decline in comparison to mental and physical well-being domains. Notably, general risk-taking propensity, which, in principle, could be related to both clusters, showed age and gender effects comparable to mental and physical well-being domains, perhaps suggesting similar interpretations of the general and well-being domains by respondents.

There are two main implications of our work and associated findings. First, our

results provide important input to theories of aging and risk taking by emphasizing the importance of domain-specific patterns that have, so far, merited relatively little theorizing. Some extant theories suggest which substantive causes may underlie differences between domains, such as the perceived costs and benefits associated with each domain or the opportunity for risk (Mishra, 2014; Weber et al., 2002). For example, reductions in physical ability with age are more likely to play a role in recreational and, perhaps, occupational domains than in other domains. Future work might study age differences in the motivations associated with engaging in different risks (Ravert et al., 2019), to shed light on the mechanisms (such as goals, costs, and opportunity) that lead to domain-specific age differences in risk-taking propensity. A similar point can be made about theories concerning gender differences in risk taking. For example, gender schema theory has so far only made general predictions about gender differences in risk taking (Frey et al., 2021) but our results suggest that it could be important to consider domain differences in future theorizing.

Second, our results have an important implication for those applications interested in assessing the role of global population aging in individual and societal risk-taking. After all, the appetite for risk is likely associated with the attitudes toward technological innovation and other consumer patterns that are central to current societal challenges (e.g., McCollum et al., 2017). As a result, modern integrated assessment models aim to include population heterogeneity in their parameters (e.g., McCollum et al., 2017; Trutnevyte et al., 2019). A quantitative assessment of age-group differences and their generality across populations can be instrumental for such efforts and our results suggest that the assumption of universal decline in the appetite for risk is warranted, which could simplify assumptions of such models that aimed at capturing age stratification in attitudes towards risk. Nevertheless, our results also indicate that significant domain and population-specific variation remains, suggesting it is likely important to consider such factors when making predictions for specific applications or populations.

Limitations and Future Directions

Our work has a number of limitations that should be acknowledged. One limitation concerns the measurement of risk-taking propensity underlying the samples selected. In many samples, risk-taking propensity was mostly captured by a single item, which likely has consequences for measurement reliability and the ability to detect systematic individual variation. Previous work has found that measures of risk-taking propensity show somewhat lower reliability over periods of decades than previously found for other major personality measures (Mata et al., 2018). In our work, we also found ICC estimates for risk-taking propensity that are somewhat lower relative to those reported for major personality traits, such as the Big Five (Graham et al., 2020), which could also indicate lower reliability of risk-taking propensity relative to other personality measures. It remains an open issue whether such results (i.e., the ratio of within- to between-subject variance) are fundamental aspects of the construct or, alternatively, result from measurement characteristics.

Secondly, one should acknowledge that our work is mostly aimed at capturing overall age differences rather than testing the impact of specific individual characteristics or life events that can account for the observed individual differences or change over time. Some past work has explicitly considered additional covariates such as individual and historical contexts to further account for time-varying differences (e.g., Dohmen et al., 2017; Jianakoplos & Bernasek, 2006; Malmendier & Nagel, 2011). We did not engage in this effort because such analyses are difficult to homogenize across many panels as these require comparable data across panels. Future work might consider selecting a set of panels for investigating the role of specific time-varying covariates, such as changes in income or marital status.

Thirdly and finally, our analyses suggest some role for period and cohort effects but our approach was relatively simple and distinguished only periods as the starting year of data collection and compared only two cohorts. Future work might consider more sophisticated methods to provide a more continuous assessment of age, period, and cohort effects on risk taking across studies and domains (Yang et al., 2021). We

hope such studies considering individual-specific differences and time-varying covariates will profit from our panel selection and publicly available code.

Summary and Conclusions

To conclude, we provide the first meta-analytic estimates of age differences in self-reported risk-taking propensity and our results suggest a systematic negative relation between age and both general and domain-specific risk-taking. Crucially, age differences are more pronounced in specific domains, such as recreational and occupational domains, relative to others, such as driving, financial, or health. Overall, our work suggests that future research is needed to clarify the underlying causes of the domain-specific nature of age differences in risk-taking propensity.

Acknowledgments

The data we used are publicly available through the original data providers and we make all our scripts and results available through our companion website (https://cdsbasel.github.io/ageriskmeta/).

This work was supported by a grant from the China Scholarship Council (CSC) to Y.L. (No.201906990029) and grants from the Swiss National Science Foundation to R.M. (http://p3.snf.ch/project-156172, https://p3.snf.ch/project-177277).

The authors thank Laura Wiles for editing the manuscript.

Author contributions

Y.L.: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, visualization, writing original draft, and writing—review & editing. A.B.: Investigation, project administration, and writing—review & editing. G.S.: Investigation and project administration. M.K.: Investigation and project administration. R.M.: Conceptualization, funding acquisition, investigation, methodology, project administration, supervision, and writing—review & editing.

Competing interests

All authors declare no competing interests related to this study.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Banks, J., Bassoli, E., & Mammi, I. (2020). Changing attitudes to risk at older ages: The role of health and other life events. *Journal of Economic Psychology*, *79*, 102208. https://doi.org/10/gmhptf

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10/gcrnkw

Best, R., & Charness, N. (2015). Age differences in the effect of framing on risky choice: A meta-analysis. *Psychology and Aging*, *30*(3), 688–698. https://doi.org/10.1037/a0039447

Bleidorn, W., Klimstra, T. A., Denissen, J. J. A., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2013). Personality maturation around the world: A cross-cultural examination of social-investment theory. *Psychological Science*, *24*(12), 2530–2540. https://doi.org/10/gf9f2q

Bonem, E. M., Ellsworth, P. C., & Gonzalez, R. (2015). Age differences in risk: Perceptions, intentions and domains. *Journal of Behavioral Decision Making*, *28*(4), 317–330. https://doi.org/10.1002/bdm.1848

Bonsang, E., & Dohmen, T. (2015). Risk attitude and cognitive aging. *Journal of Economic Behavior & Organization*, *112*, 112–126. https://doi.org/10/f6759h

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

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

Carstensen, L., Pasupathi, M., Mayr, U., & Nesselroade, J. (2000). Emotional Experience in Everyday Life across the Adult Life Span. *Journal of personality and social psychology*, *79*, 644–55. https://doi.org/10/bh47h7

Clark, W. A. V., & Lisowski, W. (2017). Prospect theory and the decision to move or stay. *Proceedings of the National Academy of Sciences*, *114*(36), E7432–E7440. https://doi.org/10/gbwzd6

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The Problem of Units and the Circumstance for POMP. *Multivariate Behavioral Research*, *34*(3), 315–346. https://doi.org/10.1207/S15327906MBR3403_2

Cross, C. P., Copping, L. T., & Campbell, A. (2011). Sex differences in impulsivity: A meta-analysis. *Psychological Bulletin*, *137*(1), 97–130. https://doi.org/10.1037/a0021591

Defoe, I. N., Dubas, J. S., Figner, B., & van Aken, M. A. G. (2015). A meta-analysis on age differences in risky decision making: Adolescents versus children and adults. *Psychological Bulletin*, *141*(1), 48–84. https://doi.org/10/f6v5m5

Del Giudice, M., Gangestad, S. W., & Kaplan, H. S. (2016). Life history theory and evolutionary psychology. In *The handbook of evolutionary psychology: Foundations, Vol. 1, 2nd ed* (pp. 88–114). John Wiley & Sons, Inc.

Depping, M. K., & Freund, A. M. (2011). Normal aging and decision making: The role of motivation. *Human Development*, *54*(6), 349–367. https://doi.org/10.1159/000334396

Dohmen, T., Falk, A., Golsteyn, B. H. H., Huffman, D., & Sunde, U. (2017). Risk attitudes across the life course. *The Economic Journal*, *127*(605), F95–F116. https://doi.org/10.1111/ecoj.12322

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

Düzel, E., Bunzeck, N., Guitart-Masip, M., & Düzel, S. (2010). NOvelty-related Motivation of Anticipation and exploration by Dopamine (NOMAD): Implications for healthy aging. *Neuroscience & Biobehavioral Reviews*, *34*(5), 660–669. https://doi.org/10/cz67sm

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

Falk, A., & Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality. *Science*, *362*(6412), eaas9899. https://doi.org/10.1126/science.aas9899

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

Frey, R., Richter, D., Schupp, J., Hertwig, R., & Mata, R. (2021). Identifying robust correlates of risk preference: A systematic approach using specification curve analysis. *Journal of Personality and Social Psychology*, *120*(2), 538–557. https://doi.org/10.1037/pspp0000287

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

Graham, E. K., Willroth, E. C., Weston, S. J., Muniz-Terrera, G., Clouston, S. A. P., Hofer, S. M., Mroczek, D. K., & Piccinin, A. M. (2022). Coordinated data analysis: Knowledge accumulation in lifespan developmental psychology. *Psychology and Aging*, *37*(1), 125–135. https://doi.org/10.1037/pag0000612

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. https://doi.org/10/dhbjj6

Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods*, *14*(2), 150–164. https://doi.org/10/fnschg

Hofer, S. M., & Piccinin, A. M. (2010). Toward an Integrative Science of Life-Span Development and Aging. *The Journals of Gerontology: Series B*, *65B*(3), 269–278. https://doi.org/10.1093/geronb/gbq017

Jianakoplos, N. A., & Bernasek, A. (2006). Financial Risk Taking by Age and Birth Cohort. *Southern Economic Journal*, *72*(4), 981–1001. https://doi.org/10.2307/20111864

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

König, A. N. (2021). Domain-specific risk attitudes and aging—A systematic review. *Journal of Behavioral Decision Making*, *34*(3), 359–378. https://doi.org/10.1002/bdm.2215

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Malmendier, U., & Nagel, S. (2011). Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?*. *The Quarterly Journal of Economics*, *126*(1), 373–416. https://doi.org/10.1093/qje/qjq004

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

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

Mata, R., Josef, A. K., Samanez-Larkin, G. R., & Hertwig, R. (2011). Age differences in risky choice: A meta-analysis. *Annals of the New York Academy of Sciences*, *1235*, 18–29. https://doi.org/10.1111/j.1749-6632.2011.06200.x

McCollum, D. L., Wilson, C., Bevione, M., Carrara, S., Edelenbosch, O. Y., Emmerling, J., Guivarch, C., Karkatsoulis, P., Keppo, I., Krey, V., Lin, Z., Broin, E. Ó., Paroussos, L., Pettifor, H., Ramea, K., Riahi, K., Sano, F., Rodriguez, B. S., & van Vuuren, D. P. (2018). Interaction of consumer preferences and climate policies in the global transition to low-carbon vehicles. *Nature Energy*, *3*(8), 664–673. https://doi.org/10.1038/s41560-018-0195-z

McCollum, D. L., Wilson, C., Pettifor, H., Ramea, K., Krey, V., Riahi, K., Bertram, C., Lin, Z., Edelenbosch, O. Y., & Fujisawa, S. (2017). Improving the behavioral realism of global integrated assessment models: An application to consumers' vehicle choices. *Transportation Research Part D: Transport and Environment*, *55*, 322–342. https://doi.org/10.1016/j.trd.2016.04.003

Mishra, S. (2014). Decision-Making Under Risk: Integrating Perspectives From Biology, Economics, and Psychology. *Personality and Social Psychology Review*, *18*(3), 280–307. https://doi.org/10.1177/1088868314530517

Piccinin, A. M., & Hofer, S. M. (2008). Integrative Analysis of Longitudinal Studies on Aging: Collaborative Research Networks, Meta-Analysis, and Optimizing Future Studies. In *Handbook of Cognitive Aging: Interdisciplinary Perspectives* (pp. 446–476). SAGE Publications, Inc. https://doi.org/10.4135/9781412976589

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

Rolison, J. J., Hanoch, Y., Wood, S., & Liu, P.-J. (2014). Risk-Taking Differences Across the Adult Life Span: A Question of Age and Domain. *The Journals of Gerontology: Series B*, *69*(6), 870–880. https://doi.org/10.1093/geronb/gbt081

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

Schurer, S. (2015). Lifecycle patterns in the socioeconomic gradient of risk preferences. *Journal of Economic Behavior & Organization*, *119*, 482–495. https://doi.org/10.1016/j.jebo.2015.09.024

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10/d9mzdb

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

Trutnevyte, E., Hirt, L. F., Bauer, N., Cherp, A., Hawkes, A., Edelenbosch, O. Y., Pedde, S., & van Vuuren, D. P. (2019). Societal Transformations in Models for Energy and Climate Policy: The Ambitious Next Step. *One Earth*, *1*(4), 423–433. https://doi.org/10.1016/j.oneear.2019.12.002

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10/gckfpj

Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263–290. https://doi.org/10.1002/bdm.414

Weston, S. J., Graham, E. K., & Piccinin, A. M. (2020). Coordinated Data Analysis: A New Method for the Study of Personality and Health. In P. L. Hill & M. Allemand (Eds.), *Personality and Healthy Aging in Adulthood: New Directions and Techniques* (pp. 75–92). Springer International Publishing. https://doi.org/10.1007/978-3-030-32053-9_6

Yang, Y. C., Walsh, C. E., Johnson, M. P., Belsky, D. W., Reason, M., Curran, P., Aiello, A. E., Chanti-Ketterl, M., & Harris, K. M. (2021). Life-course trajectories of body mass index from adolescence to old age: Racial and educational disparities. *Proceedings of the National Academy of Sciences*, *118*(17), e2020167118. https://doi.org/10.1073/pnas.2020167118

Table 1

Overview of Samples

Sample	Country	Continent	N	Female (%)	Mean age	Age range	Scale range	Survey year	Number of waves						
									G	F	D	R	O	H	S
DHS	NL	EU	9,445	46.72	52.52	18-90	1-7	1993-2020	-	29	-	-	-	-	-
GCOE Japan	JP	AS	7,014	52.32	50.96	20-77	0-10	2004-2010	7	-	-	-	-	-	-
GCOE USA	USA	NA	7,618	53.61	50.24	18-90	0-10	2005-2010	6	-	-	-	-	-	-
GLES	DE	EU	10,313	51.24	51.89	18-66	1-11	2016-2021	4	-	-	-	-	-	-
HILDA	AU	OC	20,617	48.46	50.08	18-90	1-4	2001-2019	-	16	-	-	-	-	-
HRS	USA	NA	18,614	58.52	67.27	51-90	0-10	2014-2018	4	4	4	4	4	4	-
LIKS	KG	AS	8,351	53.25	41.26	18-90	0-10	2010-2016	5	-	-	-	-	-	-
PHF	DE	EU	6,961	50.25	55.00	18-90	0-10, 1-4	2010-2016	3	3	-	-	-	-	-
SAVE	DE	EU	3,886	47.56	52.85	18-90	1-7	2001-2013	-	9	9	9	9	9	-
SHARE Austria	AT	EU	3,651	57.57	65.60	50-90	1-4	2007-2019	-	6	-	-	-	-	-
SHARE Belgium	BE	EU	3,185	53.06	64.34	50-90	1-4	2007-2019	-	6	-	-	-	-	-
SHARE Czech Republic	CZ	EU	3,452	59.73	65.43	50-90	1-4	2007-2019	-	5	-	-	-	-	-
SHARE Denmark	DK	EU	1,458	48.01	62.73	50-90	1-4	2007-2019	-	5	-	-	-	-	-
SHARE Estonia	EE	EU	5,178	61.65	66.85	50-90	1-4	2011-2019	-	3	-	-	-	-	-
SHARE France	FR	EU	3,317	55.98	65.40	50-90	1-4	2007-2019	-	5	-	-	-	-	-
SHARE Germany	DE	EU	722	50.69	65.56	50-90	1-4	2007-2019	-	6	-	-	-	-	-
SHARE Israel	IL	AS	800	49.75	68.30	50-90	1-4	2007-2019	-	3	-	-	-	-	-
SHARE Italy	IT	EU	1,913	51.75	65.56	50-90	1-4	2007-2019	-	6	-	-	-	-	-
SHARE Netherlands	NL	EU	1,598	53.94	64.37	50-90	1-4	2007-2019	-	4	-	-	-	-	-
SHARE Slovenia	SI	EU	1,853	56.77	65.90	50-90	1-4	2011-2019	-	4	-	-	-	-	-
SHARE Spain	ES	EU	2,179	55.48	66.15	50-90	1-4	2007-2019	-	5	-	-	-	-	-
SHARE Sweden	SE	EU	1,172	52.13	67.05	50-90	1-4	2007-2019	-	5	-	-	-	-	-
SHARE Switzerland	CH	EU	2,549	53.32	64.85	50-90	1-4	2007-2019	-	6	-	-	-	-	-
SOEP	DE	EU	53,608	51.57	48.95	18-90	0-10	2004-2019	14	3	3	3	3	3	3
UAS	USA	NA	7,520	59.19	49.44	18-90	0-10	2015-2021	4	-	-	-	-	-	-
USoc	UK	EU	587	56.56	51.47	18-90	0-10	2008-2014	3	-	-	-	-	-	-

* Note: For each sample, the 2-letter country code is a code set by the International Organization for Standardization (ISO) to identify each country. The continent code is a 2-letter code that identifies each continent. The survey year corresponds to the years when risk-taking propensity items were tested. G = General, F = Financial, D = Driving, R = Recreational, O = Occupational, H = Health, S = Social.

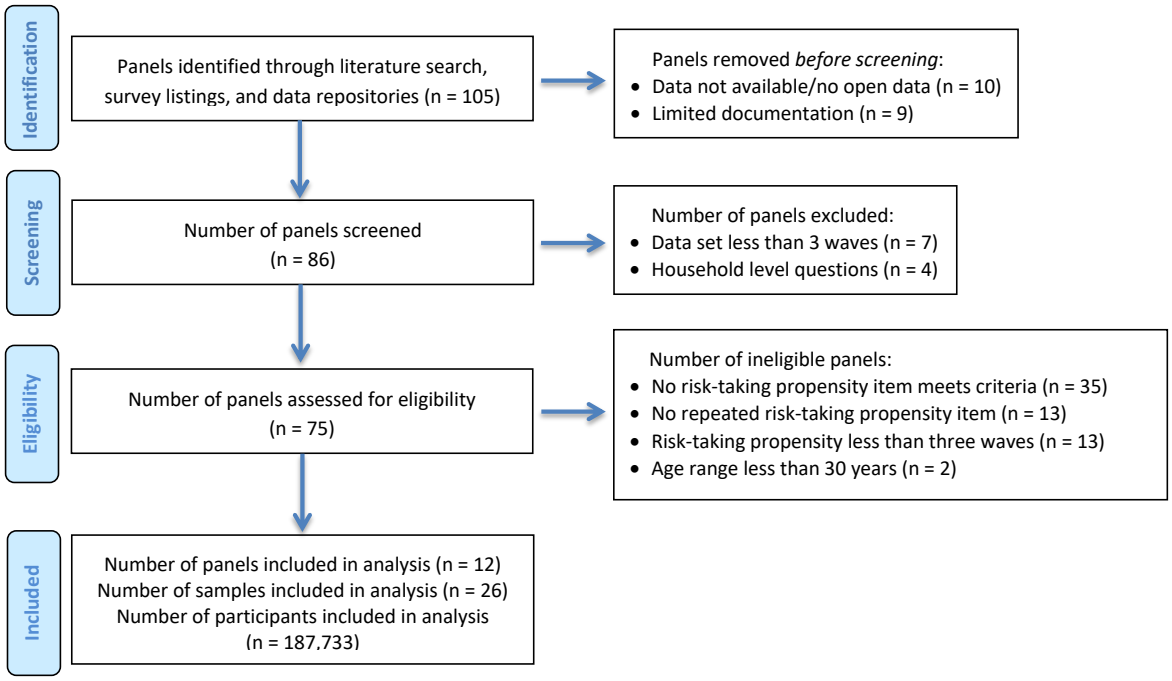


Figure 1. *Flow diagram for panel identification and selection.*

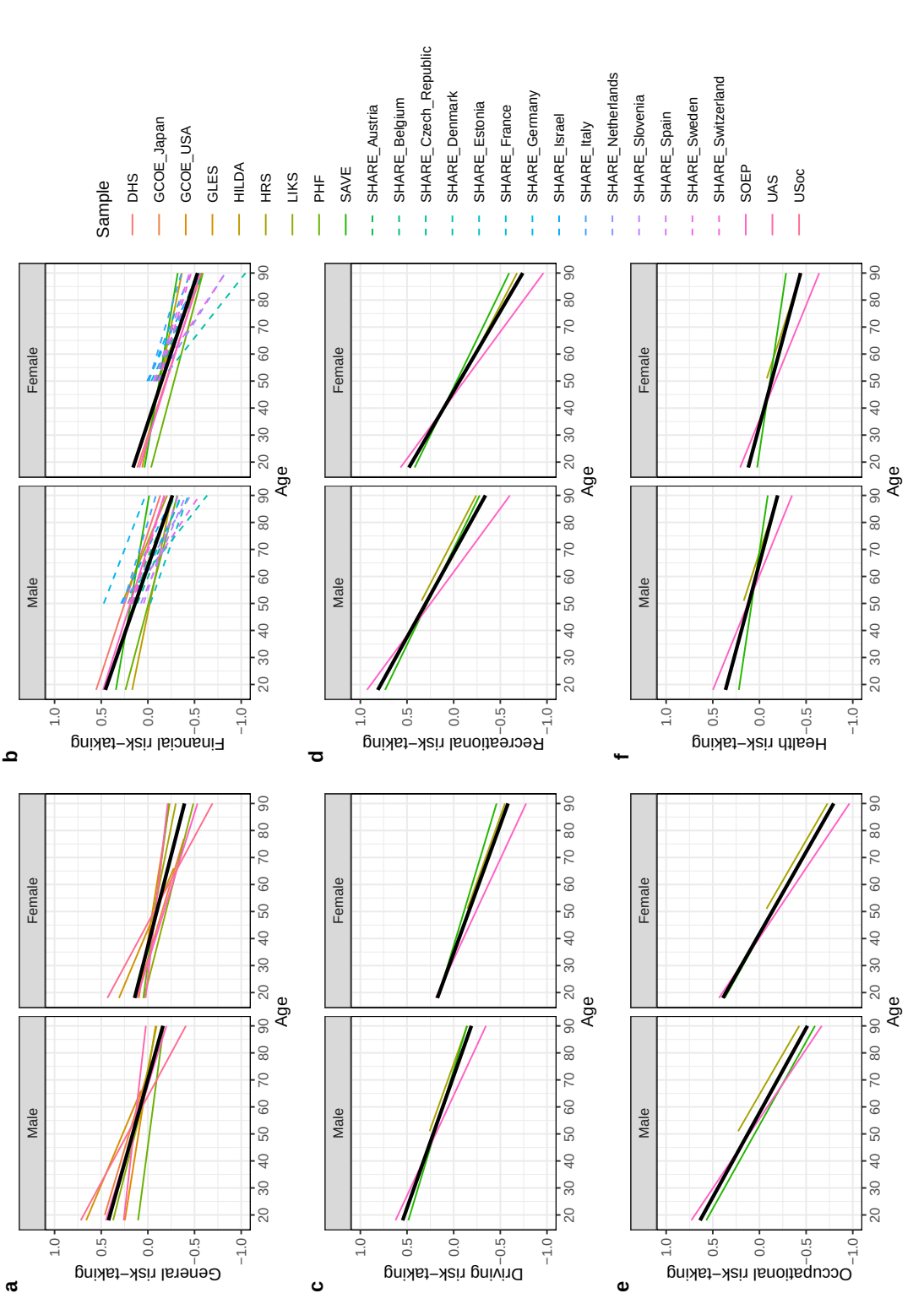


Figure 2. *Age trajectories of risk-taking propensity. Panels show results for a) general, b) financial, c) driving, d) recreational, e) occupational, and f) health domains. The solid black line indicates the average trajectory weighted by the sample size. The colored lines represent the trajectory for individual samples.*

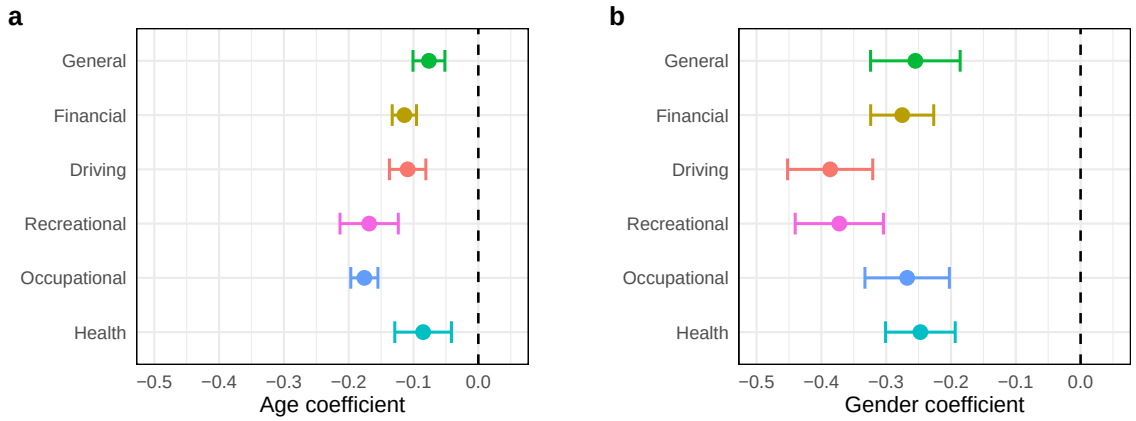


Figure 3. *Meta-analytic summary of a) the age effect per decade, and b) the gender effect on risk-taking propensity in different domains.*

Supplementary Materials

A. Panels

DNB Household Survey (DHS)

The DHS panel is a representative longitudinal panel focusing on annual financial information of Dutch households and administered by the CentERdata at Tilburg University in the Netherlands. Information and data are available via the CentERpanel platform (https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/DataPortal/DNB). Dohmen et al. (2017) previously used the 1993–2011 DHS dataset (19 waves) to estimate the age trajectory of financial risk-taking propensity. We extended this past work by including all available waves (29 waves, 1993–2021).

Preference Parameters Study of Osaka University (GCOE)

The Preference Parameters Study of Osaka University was designed to investigate economic preferences (e.g., time and risk preferences) in Japan, the United States, China (urban and rural areas), and India (urban and rural areas). Information about the panel and data access can be found at https://www.iser.osaka-u.ac.jp/survey_data/eng_application.html. For our analyses we only included data from Japan (seven waves: 2004–2010) and the USA (six waves: 2005–2010) samples as these had data on general risk-taking propensity across at least three waves.

We acknowledged that this research utilized the micro data from the Preference Parameters Study of Osaka University's 21st Century COE Program "Behavioral Macro-Dynamics Based on Surveys and Experiments", its Global COE project "Human Behavior and Socioeconomic Dynamics" and JSPS KAKENHI 15H05728 "Behavioral-Economic Analysis of Long-Run Stagnation".

German Longitudinal Election Study (GLES)

The German Longitudinal Election Study (GLES) is a project carried out in cooperation between the German Society for Electoral Research (DGfW) and Leibniz Institute for Social Sciences (GESIS). This project investigates the German political attitudes and behaviour of voters and political candidates. Detailed information can be found at the GLES homepage https://gles-en.eu/ and data are available from the GESIS Archive https://search.gesis.org/research_data/ZA6838. In our study, we included general risk-taking propensity across four waves (wave 1, 13, 14 and 15) of the GLES Panel 2016-2021 (GLES, 2021).

Household, Income and Labour Dynamics in Australia (HILDA)

The HILDA survey is a household-based panel study that collects information about the Australian population on economic and social topics. It is managed by the Melbourne Institute (Watson & Wooden, 2021). Information about the study and data access can be found at https://dataverse.ada.edu.au/dataverse/hilda. In our analyses, we included data on financial risk-taking propensity across 16 waves (2001–2019).

Health and Retirement Study (HRS)

The University of Michigan Health and Retirement Study is a longitudinal panel study of the U.S. population sponsored by the National Institute on Aging (grant number NIA U01AG009740) and managed by the Institute of Social Research, University of Michigan (Juster & Suzman, 1995; Sonnega & Weir, 2014). Information about the panel and data access can be found at https://hrs.isr.umich.edu. We analyzed general and domain-specific risk-taking propensity (financial, driving, recreational, occupational and health) across four waves (Health and Retirement Study, 2014 HRS Core, 2016 HRS Core, 2018 HRS

Core, 2020 HRS Core).

Life in Kyrgyzstan Study (LIKS)

The LIKS panel is a longitudinal survey of households and individuals in Kyrgyzstan. The panel data are available from the International Data Service Center of the Institute for Study of Labour (IDSC IZA, https://datasets.iza.org//dataset/124/life-in-kyrgyzstan-panel-study-2013). We analyzed five waves of general risk-taking propensity (2010, 2011, 2012, 2013, 2016).

Panel on Household Finances (PHF)

The PHF study is a representative and comprehensive panel survey of household finance of the German population managed by the Research Centre of the Deutsche Bundesbank (von Kalckreuth et al., 2012). Access and detailed information can be found at https://www.bundesbank.de/en/bundesbank/research/panel-on-household-finances. We analyzed general and financial risk-taking propensity across three waves (2010/2011, 2014, 2017).

We acknowledged that this paper used data from the Deutsche Bundesbank Panel on Household Finances. The results published and the related observations and analysis may not correspond to results or analysis of the data producers.

Sparen und Altersvorsorge in Deutschland (SAVE)

The SAVE panel is a representative longitudinal study of households' financial behavior managed by the Munich Center for the Economics of Aging (Coppola & Lamla, 2013). SAVE data are available from the GESIS Archive (https://dbk.gesis.org/dbksearch/GDESC2.asp?no=0014&search=save&search2=&DB=d&tab=0&notabs=&nf=1&af=&ll=10). We analyzed risk-taking propensity measures for five specific domains (financial, driving, recreational, occupational and

health) across nine waves (2001–2013).

Survey of Health, Ageing and Retirement in Europe (SHARE)

The SHARE panel aims to provide internationally comparable longitudinal data for people aged 50 or older from 28 European countries plus Israel. It is managed by the Munich Center for the Economics of Aging. Information about the panel and data access can be found at http://www.share-project.org/data-access.html. Bonsang and Dohmen (2015) used data from the first (2004) and second (2006/2007) waves of SHARE to investigate age differences in risk-taking propensity. In addition, later research also used the data from three waves (wave 2, 4 and 5) to explore risk attitudes at older ages (Banks et al., 2020). We extended previous work by analysing a total of six waves (wave 2 (2006/07), 4 (2011/12), 5 (2013), 6 (2015/16), 7 (2017/18) and 8 (2019/20), Börsch-Supan, A, 2020a, 2020b, 2020c, 2020d, 2020e, 2021), all of which included a measure of risk-taking propensity. See Börsch-Supan et al. (2013) for methodological details. Of the 29 countries included in SHARE, 14 of these (Austria, Belgium, Czech Republic, Denmark, Estonia, France, Germany, Israel, Italy, Netherlands, Slovenia, Spain, Sweden, and Switzerland) had data on a risk-taking propensity measure in at least three waves and enough respondents to conduct our analysis. For SHARE, we conducted the analyses at the country level.

We acknowledged that the SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982, DASISH: GA N°283646) and Horizon 2020 (SHARE-DEV3: GA N°676536, SHARE-COHESION: GA N°870628, SERISS: GA N°654221, SSHOC: GA N°823782) and by DG Employment, Social Affairs &

Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, and VS 2020/0313. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C, RAG052527A) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

German Socio-Economic Panel (SOEP)

The SOEP is a large multidisciplinary household survey managed by the German Institute of Economic Research, DIW Berlin (Goebel et al., 2019). Information about the panel can be found at https://www.diw.de/soep. Several studies have used the SOEP to explore age differences in risk taking. Dohmen et al. (2017) analyzed individual differences and age differences based on analysis of six waves (2004–2011) of general risk-taking propensity. Josef et al. (2016) analyzed general risk-taking across nine waves (2004–2014) and domain-specific risk-taking (financial, driving, recreational, occupational, health, and social) across three waves (2004, 2009, 2014) to study the stability and change in risk-taking propensity across adulthood. We expanded past work by using SOEP version 36 (Socio-Economic Panel (SOEP), 2021) to analyze general risk-taking propensity spanning 14 waves (2004–2019) along with domain-specific risk-taking that is available for three waves (2004, 2009, 2014), aiming to explore the age trajectory of risk-taking propensity in different domains.

Understanding America Study (UAS)

The Understanding America Study (UAS) is a nationally representative Internet panel, maintained by the Center for Economic and Social Research (CESR) at the University of Southern California (USC). The UAS survey covers multiple topics,

including cognition, personality, political views, and retirement planning. Information about the study and data access can be found at https://uasdata.usc.edu/index.php. In our research, we analysed general risk-taking data across four waves (UAS 20, UAS 95, UAS 185, and UAS 396), which also can be regarded as wave 1 to wave 4 of UAS modules corresponding topically with the HRS panel.

We acknowledged the content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of USC or UAS.

Understanding Society—the UK Household Longitudinal Study (USoc)

USoc aims to capture social and economic information about the UK's population and is managed by the Institute for Social and Economic Research (ISER) at the University of Essex. Data and information about the panel data can be accessed at https://www.understandingsociety.ac.uk/. We analyzed data from the survey's Innovation Panel (IP, University of Essex, Institute for Social and Economic Research, 2021) consisting of general risk-taking propensity for three waves (2008, 2013, 2014).

References

Banks, J., Bassoli, E., & Mammi, I. (2020). Changing attitudes to risk at older ages: The role of health and other life events. *Journal of Economic Psychology*, *79*, 102208. https://doi.org/10/gmhptf

Bonsang, E., & Dohmen, T. (2015). Risk attitude and cognitive aging. *Journal of Economic Behavior & Organization*, *112*, 112–126. https://doi.org/10/f6759h

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., Zuber, S., & on behalf of the SHARE Central Coordination Team. (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, *42* (4), 992–1001. https://doi.org/10/f5dndk

Börsch-Supan, A. (2020a). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 2. Release version: 7.1.0. SHARE-ERIC. Data set. https://doi.org/10.6103/SHARE.w2.710

Börsch-Supan, A. (2020b). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4. Release version: 7.1.0. SHARE-ERIC. Data set. https://doi.org/10.6103/SHARE.w4.710

Börsch-Supan, A. (2020c). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5. Release version: 7.1.0. SHARE-ERIC. Data set. https://doi.org/10.6103/SHARE.w5.710

Börsch-Supan, A. (2020d). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 6. Release version: 7.1.0. SHARE-ERIC. Data set. https://doi.org/10.6103/SHARE.w5.710

Börsch-Supan, A. (2020e). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 7. Release version: 7.1.1. SHARE-ERIC. Data set.

https://doi.org/10.6103/SHARE.w7.711

Börsch-Supan, A. (2021). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 8. Release version: 1.0.0. SHARE-ERIC. Data set. https://doi.org/10.6103/SHARE.w8.100

Coppola, M., & Lamla, B. (2013). Saving and Old Age Provision in Germany (SAVE): Design and Enhancements. *Schmollers Jahrbuch*, *133*, 109–116. https://doi.org/10/gmx2c8

Dohmen, T., Falk, A., Golsteyn, B. H. H., Huffman, D., & Sunde, U. (2017). Risk attitudes across the life course. *The Economic Journal*, *127* (605), F95–F116. https://doi.org/10.1111/ecoj.12322

GLES. (2021). GLES Panel 2016-2021, Waves 1-15. GESIS Data Archive, Cologne. ZA6838 Data file Version 5.0.0. https://doi.org/10.4232/1.13783

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*, *239* (2), 345–360. https://doi.org/10/gfxztr

Health and Retirement Study. (2014 HRS Core). Public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (2014).

Health and Retirement Study. (2016 HRS Core). Public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (2016).

Health and Retirement Study. (2018 HRS Core). Public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI,

(2018).

Health and Retirement Study. (2020 HRS Core). Public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (2020).

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

Juster, F. T., & Suzman, R. (1995). An Overview of the Health and Retirement Study. *The Journal of Human Resources*, *30*, S7–S56. https://doi.org/10/dqj2j7

Socio-Economic Panel (SOEP). (2021). Data for years 1984-2019, version 36, EU Edition. https://doi.org/10.5684/soep.core.v36eu

Sonnega, A., & Weir, D. (2014). The Health and Retirement Study: A Public Data Resource for Research on Aging. *Open Health Data*, *2* (1), e7. https://doi.org/10/gm2vfj

University of Essex, Institute for Social and Economic Research. (2021). Understanding Society: Innovation Panel, Waves 1-12, 2008-2019. 10th Edition. UK Data Service. SN: 6849, https://doi.org/10.5255/UKDA-SN-6849-13

von Kalckreuth, U., Eisele, M., Le Blanc, J., Schmidt, T., & Zhu, J. (2012). *The Phf: A Comprehensive Panel Survey on Household Finances and Wealth in Germany* (SSRN Scholarly Paper No. ID 2796868). Social Science Research Network. Rochester, NY.

Watson, N., & Wooden, M. (2021). The Household, Income and Labour Dynamics in Australia (HILDA) Survey. *Jahrbücher für Nationalökonomie und*

Statistik, *241* (1), 131–141. https://doi.org/10.1515/jbnst-2020-0029

Table S1

Description of models

Model	Description	Formulation (lme4 syntax)
M1	Intercept-only model	$Risk \sim 1 + (1 \mid subject)$
M2	Age fixed effects model	$Risk \sim age + (1 \mid subject)$
M3	Age fixed and random effects model	$Risk \sim age + (1 + age \mid subject)$
M4	Age fixed and random effects model with gender	$Risk \sim age + gender + (1 + age \mid subject)$
M5	Age fixed and random effects model with interaction	$Risk \sim age + gender + age \times gender + (1 + age \mid subject)$
M6	Age quadratic growth model	$Risk \sim age + age^2 + (1 + age \mid subject)$
M7	Age quadratic growth model with gender	$Risk \sim age + age^2 + gender + (1 + age \mid subject)$
M8	Age fixed and random effects model with gender and dummy cohort	$Risk \sim age + gender + age\ group + (1 + age \mid subject)$
M9	Age fixed and random effects model with gender and dummy cohort, and including interaction	$Risk \sim age + gender + age\ group + age \times gender \times age\ group + (1 + age \mid subject)$

Table S2

Meta-analytic summary of intraclass correlation coefficients from M1

Domain	Samples	ICC	*SE*	*Z*	*p*	CI.lb	CI.ub
General	9	0.46	0.04	10.48	< 0.001	0.37	0.55
Financial	20	0.36	0.03	14.23	< 0.001	0.31	0.41
Driving	3	0.47	0.04	11.78	< 0.001	0.39	0.54
Recreational	3	0.47	0.03	17.21	< 0.001	0.42	0.52
Occupational	3	0.41	0.03	11.86	< 0.001	0.34	0.48
Health	3	0.39	0.03	12.74	< 0.001	0.33	0.45

*Note: Samples = number of samples included in the meta-analysis. ICC = intraclass correlation coefficient. SE = standard error, Z = Z test, CI.ub and CI.lb = 95% confidence intervals (upper and lower bounds)

Table S3

Model comparison for samples including three or more waves

Domain	Samples	Number of samples with best-fitting		
		M3	M4	M5
General	9	1	7	1
Financial	20	0	17	3
Driving	3	0	2	1
Recreational	3	0	1	2
Occupational	3	0	3	0
Health	3	0	1	2
Social	1	0	1	0

*Note: Samples = number of samples included in the domain. M3 = age fixed and random effects model, M4 = age fixed and random effects model with gender, M5 = age fixed and random effects model with gender, including an age by gender interaction.

Table S4

Meta-analytic summary of age fixed and random effects model with gender from M4

Domain	Samples	*B*	*SE*	*Z*	*p*	CI.lb	CI.ub
Age effect							
General	9	-0.08	0.01	-6.07	< 0.001	-0.10	-0.05
Financial	20	-0.11	0.01	-11.91	< 0.001	-0.13	-0.10
Driving	3	-0.11	0.01	-7.62	< 0.001	-0.14	-0.08
Recreational	3	-0.17	0.02	-7.34	< 0.001	-0.21	-0.12
Occupational	3	-0.18	0.01	-16.47	< 0.001	-0.20	-0.16
Health	3	-0.09	0.02	-3.82	< 0.001	-0.13	-0.04
Gender effect							
General	9	-0.26	0.04	-7.23	< 0.001	-0.32	-0.19
Financial	20	-0.28	0.02	-11.09	< 0.001	-0.32	-0.23
Driving	3	-0.39	0.03	-11.53	< 0.001	-0.45	-0.32
Recreational	3	-0.37	0.03	-10.71	< 0.001	-0.44	-0.30
Occupational	3	-0.27	0.03	-8.06	< 0.001	-0.33	-0.20
Health	3	-0.25	0.03	-9.02	< 0.001	-0.30	-0.19

*Note: Samples = number of samples included in the meta-anaylsis. *B* = regression coefficient, *SE* = standard error. *Z* = *Z* test, CI.ub and CI.lb = 95% confidence intervals (upper and lower bounds).