Economic Essays on Learning, Inflation Beliefs, and Happiness

Dissertation

zur Erlangung des akademischen Grades
Doctor Rerum Politicarum

An der
Fakultät für Wirtschafts- und Sozialwissenschaften
der Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Christoph Karl Becker

Heidelberg, Mai 2023

Acknowledgements

I like to thank my supervisor Stefan Trautmann for the years of support and guidance that made me the researcher I am today. I am especially grateful for trusting in me even when I didn't.

I am deeply grateful to Christiane Schwieren, Christian Conrad and Thomas Klein, who agreed without hesitation to serve on my PhD committee. Thank you for helping me bring the research of the last years to a conclusion.

This dissertation would not have been possible (at least not without my sanity being intact) without the many friends and colleagues at the Alfed-Weber-Insitute for Economics. I like to thank in particular Freya Schadt, Christian König genannt Kersting, Martin Vollmann, Franziska Ramschütz, Robbert Schaap, Frikk Nesje, Tamás Oláh, Manuel Schick, Robert Schmidt, Pascal Kieren, Yilong Xu, Liza Charroin, Sina Klein, Marek Steinke and Alice Soldà. I am thankful for you being there for the ups and downs of research life (and many a board game evening).

I also like to extend my gratitude to other researchers from which I benefited (and hopefully returned the favor) over the years: My co-authors Eyal Ert, Gijs van de Kuilen, Tigran Melkonyan, Eugenio Proto, Andis Sofianos, Thomas Eife, Peter Dürsch, Isadora Kirchmaier and Alexander Glas. Nick Chater and Hossam Zeitoun, who graciously hosted my visit to Warwick Business School, and my fellow PhD students and researchers there, Fendi Tsim, Rafael Alejandro Vaquera Salazar, Archit Tamboli, Junyang Guo, Christina Kataroushi and Tahani Almabadi. Furthermore, I am obliged to Patricia von Mellenthin for further scrutinizing this thesis and Robin Schiel for helping to ensure that it gets finished.

Finally, I like to thank my friends and family back home. My family has no academic background and I appreciate my parents supporting my chosen path. My family has always been there to provide a desperately needed sense of normalcy when my things turned crazy. My deepest thanks also go out to my friends, Thomas Rietbrock, Stefan Rietbrock, Lina Duarte, Michael Ganninger, Jessica Schuster and Jonathan Fröhlich.

Contents

Introduction

"The economists' only hopeful objective is to provide an account of that shawl of loosely interknitted strands which waves in the wind of other human influences (...)"
– George L. S. Shackle; Epistemics & Economics 2017, page 240

"Science is riddled with stories. In fact, if you can't tell a convincing story about your research, nobody will let you publish it."
– Terry Pratchett, Ian Stewart and Jack Cohen; The Science of Discworld II 2011, page 326

Economics can be contrasted from other social sciences by viewing human affairs primarily as allocation problems, where only a finite stock of resources is available (Mankiw, 2020). Be it natural resources, energy, money, time, or attention: In each of these situations, it is often challenging to decide how to optimally allocate these resources. Economics can help to understand how people make these decisions, what pitfalls should be avoided, and how to create mechanisms for consistent decisions. To that end, this dissertation contains essays covering three aspects of human affairs: Learning, inflation expectations, and happiness. While these three topics strongly differ from each other, all studies presented here have a common denominator: They concern beliefs, assessments, and expectations of individual human beings. These in turn provide a basis for individual actions that sum up to lead to the phenomena studied in macroeconomics or empirical research. Thus, the studies presented here, while differing widely, examine some of the driving forces behind economic activity.

As in any other scientific discipline, multiple approaches can be used to gather knowledge. The studies within this dissertation approach the three aspects by experimental and empirical methods. Using data from laboratory experiments, I first provide insights into how learning shapes risk preferences and how decision makers gather information about previously unforeseen events. While laboratory settings abstract from real-world decision making, they provide the opportunity to study behavior in a tightly controlled setting. Even though the results from these experiments cannot easily be extrapolated to settings outside the laboratory, they point towards the fundamental mechanisms of behavior that warrant further investigation.

Inflation is a much more specific phenomenon and, especially while writing this in 2023, a very relevant one for everyday purchases. Therefore, it might be detrimental to study inflation beliefs solely in the abstract. The two studies on inflation beliefs bridge the gap between laboratory and empirical research. The first study uses data gathered via Prolific, an online subject pool. While the overall setup in this study is still reminiscent of laboratory experiments, it provides a much broader cross section of the overall population than typical convenience samples of students do. For the second study, we were able to introduce treatments from our experiment into the online household panel survey of the German Bundesbank, giving us a representative sample of the German population.

For happiness, finally, the studies presented focus mainly on the interaction between age, social networks, and subjective measures of happiness. Therefore, the interest of these studies is not on momentary effects of a specific intervention on happiness, as one could measure in the laboratory, but on longer-term trends. To that end, both studies employ empirical methods using data from the Survey of Health, Ageing and Retirement in Europe (SHARE, Börsch-Supan et al. 2013).

In the following paragraphs, I summarize my work on each topic by giving a brief overview of each study included in my thesis.

Learning

Learning is a fundamental human experience. From getting taught in kindergarten and school, to trying out obtained knowledge in new circumstances, or to making completely new experiences, learning is ever-present. As such, it is of fundamental interest to the behavioral sciences. Models or games in economics in the past often involved situations in which all relevant knowledge is readily available to decision makers. In the last decades the shortcomings of many economic models led economists to look more into the other behavioral sciences and especially psychology, where learning has been studied intensively. In the field, studying learning can be complicated, as it is difficult to retain control over what decision makers also learn outside of the task at hand. Studies have thus tried to design experiments that artificially recreate a learning experience in the laboratory. These artificial learning simulations usually require decision makers to *sample* information from different options, often involving lotteries offering different rewards (see Hertwig & Erev, 2009 for a concise overview of different paradigms used in the laboratory).

Chapter 1 examines how certain risk preferences are influenced by having to sample relevant information rather than receiving a full description. As already pointed out by Knight (1921), decision makers need to differentiate between known *a-priori* probabilities (e.g., the probability of tails in a fair coin toss) and *statistical* probabilities, which are derived empirically and might not accurately reflect the underlying distribution. Accordingly, numerous studies have found a description-experience gap: Elicited risk attitudes differ if decision makers (i) have information about all relevant probabilities and outcomes or (*decisions from description*) (ii) if this information must be obtained through experience sampling (*decision from experience*, Wulff et al., 2018; Hertwig & Wulff, 2022). Specifically, decisions from description tend to be risk averse for large-probability gains and risk seeking for small-probability gains, potentially explaining why people shun stock investments but play lotteries. In decisions from experience, this pattern reverses (Hertwig, 2012).

However, studies in economics and finance in the meantime have pointed out that looking at the degree of risk aversion (second degree risk attitude) alone is often insufficient to explain how decision makers choose (Noussair et al., 2014). Instead, these studies advocate to look more to the so-called *higher-order risk attitudes*, often *prudence* (third degree risk attitude, conceptually related to the third derivative of a utility function) and *temperance* (fourth degree risk attitude, conceptually related to the fourth derivative of a utility function) (Eeckhoudt & Schlesinger, 2006). While the technical definitions in the expected utility framework are a bit more complex (Ebert & Wiesen, 2011), these preferences are often approximated by a decision maker's preference concerning the *skewness* and the *kurtosis* of a distribution. These measures provide insight into the symmetry and "tailedness" of a distribution and thus capture more than just looking at the variability of outcomes. Especially prudence/skewness seeking has received strong attention, as prudent decision makers are expected to build up precautionary savings when faced with additional risks, which tracks with empirical findings (Noussair et al., 2014).

Results on higher-order risk attitudes are again mainly derived from decisions from description. This might not accurately reflect the environment in which decision makers make decisions in real-life settings, such as repeatedly making investment or saving decisions. In Chapter 1, my coauthors and I study how these higher-order risk attitudes are affected by learning in the abstract. Participants make decisions both from description and experience between two possibly payoff-relevant lotteries. For the experience tasks, participants have no information about the two lotteries but are free to sample information from them. Across two experiments, we find that participants on average do not exhibit temperance. We also find that participants are on average prudent, in line with previous literature (Trautmann & van de Kuilen, 2018), but they exhibit significantly less prudence when they need to acquire information via sampling. These results hold whether the number of samples they can draw is fixed to 20 (experiment 1) or whether it can be flexibly chosen by them (experiment 2). Additionally, participants in experiment 2 only draw a relatively small number of samples, that is roughly comparable to the numbers of samples allowed in experiment 1. As a result of this, participants tend to undersample the lotteries, which might bias their impression of their variance, skewness, and kurtosis. Indeed, we find that participants that sample more exhaustively in experiment 2, also exhibit more stable preferences across the description and experience tasks. Finally, we replicate a common finding from the literature, namely that participants in sampling tasks tend to be more expected value maximizing (Erev et al., 2010).

Chapter 2 concerns *unforeseen* and *unforeseeable* events which might fundamentally change our model of reality. Real-life examples of such events include global pandemics, political and economic crises, as well as groundbreaking scientific discoveries. While some might be aware of the chance of some of such events, the average decision maker might simply fail to account for them. This poses the question if we can isolate how decision makers react when they are faced with an event that they might previously have not considered at all. In the two experiments in the second study, we use a similar approach as in the previous studies: Participants are confronted in each choice tasks with one or two lotteries presented as urns containing marbles that represent different outcomes. They then either observe a sample (experiment 1) or draw an individual sample themselves from the urn (experiment 2). In experiment 1, unforeseen events are introduced at one point by adding another urn containing new outcomes to the original urn. Experiment 2, on the other hand, was intended to study the evolution of beliefs over time. Here the unforeseen events are conceptualized as discovering a new outcome for the first time in the urn.

A problem when studying unforeseen events is that classical learning models based on Bayesian updating are silent about how a decision maker should update, as an unforeseen event has no prior. This makes hypothesizing about how a decision maker reacts difficult. *Reverse Bayesianism* (Karni & Vierø, 2013; Karni et al., 2020) attempts to close this gap by allowing decision makers to flexibly choose the probability estimate for a newly discovered outcome while imposing constraints on estimates for already known outcomes. Specifically, reverse Bayesianism requires that the ratios of estimates of already known events remain constant to each other, even though their individual estimates might be changed. This ensures that the new model of reality is consistent with the old one. In practice this corresponds to applying Bayes' law in reverse to the estimate of the new outcome (going back in time, as if we learned that the outcome is not possible). Participants in our experiments comply with this reasoning, even when the urn contains up to four outcomes. This is noteworthy, as the number of ratios that need to be kept constant increases with each observed outcome, making the task more difficult. This holds despite participants also exhibiting other known violations of classical Bayesian updating. Specifically, we find evidence for *unpacking bias/partition dependence* (Tversky & Koehler, 1994; Fox & Rottenstreich, 2003), where participants increase the belief assigned to a hypothesis (e.g., chance of bad weather tomorrow), after it is split into smaller hypotheses (e.g., chance of rain tomorrow and chance of hail tomorrow).

Inflation beliefs

The next two chapters concern the measurement of individual inflation beliefs. Central banks take a strong interest into measuring these beliefs in order to inform and gauge the effect of their monetary policy, especially that of policies such as forward guidance. An increasingly common way to elicit inflation beliefs is using *probabilistic/density forecasts*, in which respondents assign probabilities to pre-defined ranges of inflation (Armantier et al., 2017). Numerous surveys, such as the New York Fed's Survey of Consumer Expectations (SCE) or the Bundesbank Online Panel for Households (BOP-HH), have already adopted such forecasts. For ostensibly good reasons: Density forecasts give researchers a full distribution of beliefs and allow to not only measure the average inflation respondents expect, but also to derive additional measures, such as forecast uncertainty (De Bruin et al., 2011). One problem of this method is, however, that beliefs stated in such questions appear to be sensitive to the response scale used (Schwarz, 2010). Modifying the response scale, for example by shifting the center of it, also seem to modify beliefs. This poses the question how reliable beliefs obtained in density forecasts actually are.

In Chapter 3, my coauthors and I aim to shed light on this question by introducing systematic modifications to the scale of the SCE density forecast. These modifications include shifting the entire scale, expanding, or shrinking it, as well as adding or dividing intervals. We find that while respondents tend to follow our treatment interventions (i.e., shifting the center of their distribution if the entire scale is shifted), there are limits to this effect. For example, respondents do not shift their beliefs strongly towards deflation, if the entire response scale is shifted into that direction. We take this as an indication that respondents are not simply just biased, but at least partly take the response scale provided as information and use it to update their own priors. In addition, we find that modifying the response scale can have further consequences. For example, we find that respondents allocate larger probability mass to ranges of inflation with more intervals on the response scale. This is in line with the concept of partition dependence, as discussed above, and can influence any measure derived from density forecasts. Overall, the technical and precise appearance of density forecasts might be somewhat misleading. Accordingly, more care should be taken when designing and using probabilistic questions. At the end of our study, we discuss some possible amendments. Simply canceling out a possible measurement bias introduced by the scale by taking differences over time might not be feasible, as the bias itself might vary over time. Instead, we recommend changes to how density forecasts are elicited. Using response scales with equal-width intervals and anchoring the scale to a previously elicited point forecast seem to be promising improvements that are easy to implement.

Chapter 4 provides further insight into this suggestion. As a spin-off from the previous study, my coauthors and I were able to add two of our treatment variations to the BOP-HH, which uses the same question format as the SCE. The data was collected by the Bundesbank in June 2022, a point in time when not only the inflation rates were higher, but also when the discussion around inflation had increasingly entered the public discourse. We find that the heightened inflation rates have two major consequences for respondents trying to reflect this in their beliefs in the original Bundesbank density forecast: First, the number of intervals on the response scale is higher around the center. Respondents wanting to express higher inflation beliefs need to do so with a smaller number of intervals at their disposal. Second, the response scale is open at both ends. Specifically, the last inflation interval covers all values from inflation rates of 12% or higher (for comparison, the inflation rate in June 2022 was already 7.9%). Probability mass allocated to this open interval comes with a loss of information, as we do not know the maximum inflation rate the respondent considers. Indeed, we find that respondents in the original question overall use less intervals and allocate more probability mass to the rightmost open interval. The data obtained is coarser as a result.

Using instead one of our shift treatments, where the entire response scale is shifted by 4 percentage points, already remedies this problem in part. Respondents in this treatment on average use more intervals and allocate substantially less probability mass to the open interval than those in the original format. Importantly, this also increases the consistency of the mean derived from the density forecast with the inflation point forecast, elicited in the same survey. Using simulations, we demonstrate that this result is not caused by unresponsive respondents simply transcribing unchanged beliefs to a different response scale. We also find tentative evidence that our treatment intervention in part also carries over to the subsequent wave of data collection. Taken together, these results highlight the importance of adapting the response scales in times of high inflation. While care has to be taken when changing density forecasts, as also demonstrated by the first study, fundamental changes in macroeconomic conditions need to be reflected in question design.

Happiness

Finally, the last two chapters focus on measuring individual happiness. Economists have long been engaged in measuring the welfare and prosperity of societies, mainly for practical applications: How well does the economy of a society perform? In which areas should additional funds be invested? In the past, the prime economic measure of welfare was the gross domestic product (GDP). However, as often criticized, the GDP does not measure crucial elements of the economy such as care work, housework, or other forms of unpaid work. This was already recognized by Simon Kuznets, one of the major contributors to the measurement of GDP: "The welfare of a nation can scarcely be inferred from a measure of national income" (Kohler & Chaves, 2003). Different approaches have aimed to reconcile these problems by measuring welfare more broadly. Two of the most prolific examples are the UN's Human Development Index (HDI), suggested and developed by economists such as Mhabub ul Haq and Amartya Sen, and the Happy Planet Index (HPI). Another approach is to measure the individual happiness of members of a society (Deaton, 2008). This can be done in multiple ways, the most common methods being self-reported happiness or life-satisfaction, measures of positive and negative affect, or indirect measures, such as the number of antidepressants consumed.

There have been numerous debates around alternative measures of welfare and particularly the term happiness. One common critique is that happiness, in the sense of being in a state of bliss, might not always be a desirable goal. Rather, satisfaction can also be found in struggling for a worthwhile purpose or living a live according to one's philosophical principles (cf. *Eudaimonia*, see Aristoteles' Nicomachean Ethics, e.g., in Ameriks & Clarke 2000). I fully agree with this critique. However, it is difficult to find a fitting catch-all term for the varying measurements outlined here. In the literature, the term *happiness* is often used as a shorthand for any approach trying to measure well-being, fulfillment, or forms of happiness that capture more than just material saturation. Hence, in the remainder of this thesis I will use the term happiness in this manner. In the two studies concerning happiness, in total four measures for happiness are used: Self-reported well-being; the CASP-12, which measures quality of live in terms of control and autonomy; a self-report of depressive symptoms; and satisfaction with one's social network. These studies, while summed up simply as studies on happiness, therefore measure a larger array of different aspects of welfare.

As mentioned, both studies use empirical data from the Survey of Happiness, Ageing and Retirement in Europe (SHARE). This data allows us access to thousands of households and individuals across many European countries (including Israel) and over multiple years. SHARE is mainly designed to track the life situation of older Europeans towards and past retirement. As it also includes the above-mentioned happiness measures, it also allows us to track how happiness involves for older Europeans.

Chapter 5 looks at social networks and their impact on happiness. We define social networks as "a network of family, friends, neighbors, and community members that is available in times of need to give psychological, physical, and financial help" (NCI Dictionary of Cancer Terms, 2018). Past literature is divided how exactly family status and parenthood contribute to happiness (e.g., Mastekaasa, 1994; Hansen, 2012; Litwin & Stoeckel, 2013, 2014). We study the effect different network types (e.g., family, friends, or spouse networks) and find that social networks in general have a positive effect. This positive effect is roughly comparable across networks. An important implication of this is that social networks could be substituted, i.e., that the loss of friends and family could successfully be compensated. As a side result, we also report that the direct effect of children is negative, but positive after they leave the household. A possible explanation here might be that younger children require more care and thus lead to increased stress, but that this effect might lessen the more mature and self-reliant they become. Leaving the household might simply be a good proxy for this underlying effect.

Chapter 6 takes the SHARE data and relates it to the discussion around the so-called *U-shape of happiness*. There has been a vast amount of literature reporting that happiness follows a U-shape over life: That is, happiness decreases until middle age and then starts to increase again as people grow older (Blanchflower, 2021). This U-shape can also be obtained when looking at indirect measures of happiness, e.g., the use of antidepressants (Giuntella et al., 2023). On the other hand, there also have been numerous papers arguing that such a U-shape is an artifact of using incorrect statistical methods (e.g., Glenn 2009; Frijters & Beatton 2012). A subset of these studies also points out that while the U-shape might exist, the term U-shape is incorrect when considering that there might be an additional decrease for very old respondents (Laaksonen, 2018). The study presented here is intended to remedy some of these issues, using SHARE not only as a large data set, but also as one that contains data across countries and time. We include specifications with different control sets (separating between exogenous factors such as sex and more endogenous ones, such as income), control for cohort effects, use different specifications and control for attrition (as those unhappier might simply die earlier). As a caveat, we can only test if happiness increases after middle age, as SHARE only includes data for older Europeans. Therefore, we test if the second half of the U-shape emerges in SHARE data.

Overall, we find support for the second half of the U-shape, that is an increase in happiness past middle age. This finding holds no matter the controls used and irrespective of the concrete regression model used. While we find attrition effects, i.e., that unhappier respondents tend to die earlier, the pattern still holds for the subsample of respondents that where present in all waves of our panel. Furthermore, we also find that happiness tends to decrease for the very old, possibly as health starts to deteriorate and personal losses might mount up. The pattern is present for both men and women, but less pronounced for women. Looking at individual countries, the pattern cannot always be obtained. However, as the sample size varies considerably between different countries, we cannot provide an answer if this is due to the sample size or due to fundamental differences in the age-happiness relation across countries.

Introduction References

Ameriks, K., & Clarke, D. M. (2000). *Aristotle: Nicomachean Ethics*. Cambridge University Press.

Armantier, O., Topa, G., Van der Klaauw, W., & Zafar, B. (2017). An overview of the survey of consumer expectations. *Economic Policy Review*(23-2), 51–72.

Blanchflower, D. G. (2021). Is happiness U-shaped everywhere? Age and subjective well-being in 145 countries. *Journal of Population Economics*, *34*(2), 575–624.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., ... Zuber, S. (2013, August). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, *42*(4), 992–1001.

Deaton, A. (2008). Income, health, and well-being around the world: Evidence from the Gallup World Poll. *Journal of Economic Perspectives*, *22*(2), 53–72.

De Bruin, W. B., Manski, C. F., Topa, G., & Van Der Klaauw, W. (2011). Measuring consumer uncertainty about future inflation. *Journal of Applied Econometrics*, *26*(3), 454–478.

Ebert, S., & Wiesen, D. (2011). Testing for prudence and skewness seeking. *Management Science*, *57*(7), 1334–1349.

Eeckhoudt, L., & Schlesinger, H. (2006). Putting risk in its proper place. *The American Economic Review*, *96*(1), 280–289.

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

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, *14*(3), 195–200. (Publisher: SAGE Publications Sage CA: Los Angeles, CA)

Frijters, P., & Beatton, T. (2012). The mystery of the U-shaped relationship between happiness and age. *Journal of Economic Behavior & Organization*, *82*(2-3), 525–542.

Giuntella, O., McManus, S., Mujcic, R., Oswald, A. J., Powdthavee, N., & Tohamy, A. (2023). The midlife crisis. *Economica*, *90*(357), 65–110. (Publisher: Wiley Online Library)

Glenn, N. (2009). Is the apparent U-shape of well-being over the life course a result of inappropriate use of control variables? A commentary on Blanchflower and Oswald (66: 8, 2008, 1733–1749). *Social Science & Medicine*, *69*(4), 481–485.

Hansen, T. (2012). Parenthood and Happiness: a Review of Folk Theories Versus Empirical Evidence. *Social Indicators Research*, *108*(1), 29–64.

Hertwig, R. (2012). The psychology and rationality of decisions from experience. *Synthese*, *187*(1), 269–292. (Publisher: Springer)

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523.

Hertwig, R., & Wulff, D. U. (2022). A description–experience framework of the psychology of risk. *Perspectives on psychological science*, *17*(3), 631–651.

Karni, E., Valenzuela-Stookey, Q., & Vierø, M.-L. (2020). Reverse bayesianism: A generalization. *The BE Journal of Theoretical Economics*, *21*(2), 557–569.

Karni, E., & Vierø, M.-L. (2013). "Reverse Bayesianism": A choice-based theory of growing awareness. *American Economic Review*, *103*(7), 2790–2810.

Knight, F. H. (1921). Risk, uncertainty and profit. *New York: Hart, Schaffner and Marx*.

Kohler, G., & Chaves, E. J. (2003). *Globalization: critical perspectives.* Nova Publishers.

Laaksonen, S. (2018). A research note: Happiness by age is more complex than U-shaped. *Journal of Happiness Studies*, *19*(2), 471–482.

Litwin, H., & Stoeckel, K. J. (2013). Social networks and subjective wellbeing among older Europeans: does age make a difference? *Ageing and Society*, *33*(07), 1263–1281.

Litwin, H., & Stoeckel, K. J. (2014, October). Confidant Network Types and Well-Being Among Older Europeans. *The Gerontologist*, *54*(5), 762–772.

Mankiw, N. G. (2020). *Principles of economics.* Cengage Learning.

Mastekaasa, A. (1994). Marital Status, Distress, and Well-Being: An International Comparison. *Journal of Comparative Family Studies*, *25*(2), 183–205.

NCI Dictionary of Cancer Terms, N. C. I. I. (2018, October). *Definition of social support.* Retrieved from `https://web.archive.org/web/20181023124317/https://www.cancer.gov/publications/dictionaries/cancer-terms/def/social-support`

Noussair, C. N., Trautmann, S. T., & Van de Kuilen, G. (2014). Higher order risk attitudes, demographics, and financial decisions. *Review of Economic Studies*, *81*(1), 325–355.

Schwarz, N. (2010). Measurement as cooperative communication: What research participants learn from questionnaires. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The sage handbook of measurement* (pp. 43–61). Los Angeles: Sage.

Shackle, G. L. S. (2017). *Epistemics and economics: A critique of economic doctrines.* Routledge.

Stewart, I., Cohen, J., & Pratchett, T. (2011). *The science of discworld ii: The globe.* Random House.

Trautmann, S. T., & van de Kuilen, G. (2018). Higher order risk attitudes: A review of experimental evidence. *European Economic Review*, *103*, 108–124.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547.

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin*, *144*(2), 140.

Chapter 1

Experiencing Risk: Higher-order Risk Attitudes in Description- and Experience-based Decisions

Abstract[†]

Risky decisions are often characterized by (a) imprecision about consequences and their likelihoods that can be reduced by information collection, and by (b) unavoidable background risk. This paper addresses both aspects by eliciting risk attitude, prudence and temperance in decisions from description and decisions from experience. The results reveal a novel description-experience gap for prudence, and replicate the known gap for risky decisions. While widespread prudence has been observed in decisions form description, we find no evidence of prudent decision making from experience. In decisions from experience people are strongly influenced by the sampled mean, while skewness plays a smaller role than in decisions from description.

[†]Joint work with Eyal Ert, Stefan T. Trautmann and Gijs van de Kuilen.
The content of this chapter has been published as: Becker CK, Ert E, Trautmann ST, van de Kuilen G. Experiencing risk: Higher-order risk attitudes in description- and experience-based decisions. J Exp Psychol Learn Mem Cogn. 2021 May;47(5):727-746.

1.1 Introduction

Uncertainty is a common component of many decisions, and people often try to avoid or reduce known uncertainties. These tendencies are typically quantified with measures of risk aversion. Avoiding a risk is, however, not always possible. Many decisions involve unavoidable background risks, either due to the specific environment (e.g. inflation for workers bargaining about wages), or due to other peoples' decisions (e.g., while driving on a crowded highway). In these cases, measures of risk aversion are often insufficient to explain observed decisions, both theoretically and empirically (Noussair et al., 2014; White, 2008). Analyses of decision making need to consider higher-order risk attitudes in these cases, in particular prudence and temperance (Eeckhoudt & Kimball, 1992).

Broadly speaking, prudence is the tendency towards precautionary behavior. When faced with an unavoidable risk affecting future revenues, a prudent person strives to take precautionary measures against this risk. For example, when faced with the risk of losing current employment due to an economic collapse, a prudent person builds additional savings as a buffer (Eeckhoudt & Schlesinger, 2006; Kimball, 1990). Equivalently, a prudent person is averse to an increase in downside risk (more left-skewed risk, Menezes et al. 1980), respectively has preference for positive skewness (Mao, 1970; Eeckhoudt & Schlesinger, 2006; Ebert, 2013), and exhibits a positive third degree derivative in the expected utility framework (Kimball, 1990; Eeckhoudt & Schlesinger, 2006). Temperance refers to the tendency to disaggregate multiple risks if possible (Eeckhoudt & Schlesinger, 2006; Ebert, 2013), not putting "all the eggs into one basket." Temperance can equivalently defined as an individual exhibiting a negative fourth degree derivative in the expected utility framework (Kimball, 1990; Ebert, 2013), or as being averse to distributions with high kurtosis (Eeckhoudt & Schlesinger, 2006).

Seminal work by Eeckhoudt & Schlesinger (2006) showed that prudence and temperance in terms of (unobservable) utility properties under expected utility is equivalent to an (observable) preference between lotteries in a simple binary choice framework. That is, a decision maker revealing the proposed preference between lotteries must be prudent (resp. temperate) under expected utility. Moving beyond expected utility, however, Eeckhoudt & Schlesinger (2006) use the simple lottery choices to define a purely behavioral measure of prudence and temperance (independently of expected utility assumptions). We employ these measures in the current study. In particular, let $x > y$ be two monetary outcomes, and let z_1 and z_2 be two zero-mean risks, then Eeckhoudt & Schlesinger (2006) define prudence as a preference for lottery $(0.5 : x + z_1; 0.5 : y)$ over $(0.5 : x; 0.5 : y + z_1)$, and temperance as a preference for lottery $(0.5 : x + z_1; 0.5 : x + z_2)$ over $(0.5 : x + z_1 + z_2; 0.5 : x)$. In our study, we parametrize the prudence measure for example with $x = 10, y = 6$, and $z_1 = (0.75 : 2; 0.25 : -6)$ in choice Prudence 1 (see Table 1.1, Set 1), and temperance for example with $x = 9, z_1 = (0.5 : 3; 0.5 - 3)$ and $z_2 = (0.5 : 3; 0.5 - 3)$ in choice Temperance 1 (see Table 1.1, Set 1). Studies employing this method found that especially prudence is a widespread attitude, while temperance is less prevalent (Ebert & Wiesen, 2014; Noussair et al., 2014; Trautmann & van de Kuilen, 2018). Importantly, behavioral measures of higher order risk attitude are a good predictor of people's behavior in decisions with unavoidable background risk across a range of domains from finance to health behavior (Noussair et al., 2014; Schneider & Sutter, 2020). That is, these studies suggest the validity of lottery-choice measures as an instrument of the underlying psychological motives.

Table 1.1: Decision tasks.

Task	**Set 1**			**Set 2**		
	Risky, imprudent or intemperate		Safe, prudent or temperate	Risky, imprudent or intemperate		Safe, prudent or temperate
Risk aversion task 1 (high prob.)	0.9 - €15 0.1 - €0	vs.	€13.5	0.9 - €15.5 0.1 - €0.5	vs.	€14
Risk aversion task 2 (low prob.)	0.1 - €15 0.9 - €0	vs.	€1.5	0.1 - €15.5 0.9 - €0.5	vs.	€2
Prudence task 1	0.5 - €10 0.375 - €8 0.125 - €0	vs.	0.375 - €12 0.5 - €6 0.125 - €4	0.5 - €10.5 0.375 - €8.5 0.125 - €0.5	vs.	0.375 - €12.5 0.5 - €6.5 0.125 - €4.5
Prudence task 2 (non ES)	0.5 - €9 0.4 - €7.5 0.1 - €0	vs.	0.4 - €11.5 0.5 - €6 0.1 - €3	0.5 - €9.5 0.4 - €8 0.1 - €0.5	vs.	0.4 - €12 0.5 - €6.5 0.1 - €3.5
Temperance task 1	0.125 - €15 0.75 - €9 0.125 - €3	vs.	0.5 - €12 0.5 - €6	0.125 €15.5 0.75 - €9.5 0.125 - €3.5	vs.	0.5 - €12.5 0.5 - €6.5
Temperance task 2	0.125 - €20 0.75 - €10 0.125 - €0	vs.	0.5 - €15 0.5 - €5	0.125 €20.5 0.75 - €10.5 0.125 - €0.5	vs.	0.5 - €15.5 0.5 - €5.5

The way decision makers treat uncertainty depends also on the information about possible outcomes and their probabilities. According to Knight (1921), decision makers need to differentiate between a-priori probabilities (e.g. the probability of tails in a fair coin toss) and statistical probabilities, which are derived empirically. Outside of a small set of situations like the mentioned coin toss or casino gambling, most decisions fall into the second category. While there are different ways to infer probabilities empirically, the most straightforward one is to rely on past experiences. Every time a decision maker observes the consequences of a decision, new information is acquired that can be used in the next instance a similar decision situation occurs. It is well known that decision makers' degree of risk aversion depends on whether information about a prospect is readily available, or whether it needs to be collected through experience and sampling. The difference between the two information formats is called the description-experience gap (Erev et al., 2010; Hau et al., 2010; Hertwig & Erev, 2009; Wulff et al., 2018). For example, the so-called fourfold pattern of risk attitudes[1] reverses if there is no a-priori information and sampling is required. That is, decision making processes differ substantially depending on whether risk is described, or experienced by the decision maker. In a recent meta-analysis, Wulff et al. (2018) summarize the main results of the experience-sampling paradigm: i) decision makers rely on small sets of experience, ii) there are no strong recency effects, iii) sampling error cannot fully explain behavior, and iv) decision makers tend to make decisions more in line with expected value maximization (compared to the description paradigm).[2]

Most risky decisions outside the decision analyst's lab take place in a context close to the decision from experience paradigm and concern trade-offs in the presence of unavoidable background risk. In such environments, higher order risk attitudes are predicted to matter most from a decision theoretic point of view (Brunnermeier et al., 2007; Ebert et al., 2017; Eeckhoudt & Gollier, 2005; Eeckhoudt & Kimball, 1992). Moreover, the presence of tail events which are reflected by the higher order moments of risky distributions, notably their skewness, has important implications for risk behavior in financial markets (de Roon & Karehnke, 2018; Harvey & Siddique, 2000). Higher order risk attitudes relate closely to these higher order moments. However, financial market outcomes are naturally perceived by most investors in an experience-based way, reducing the ecological validity of decision-from-description assessments of higher order risk attitude for these markets.

Previous literature emphasized the importance of sampling error (Fox & Hadar, 2006) and underweighting (Hertwig & Erev, 2009; Ungemach et al., 2009) in decisions from experience. Both aspects are important in distributions with tail events, like those to determine a decision makers prudence and temperance: Tail events may be under-sampled and under-weighed at the same time. Assessing the prevalence of prudence and temperance in situations with experience and sampling is therefore warranted. Despite the recent wave of studies assessing higher order risk attitudes in various setups and subject pools, there is scarce evidence on their role within the decisions from experience paradigm, and no evidence on the underlying decision processes. Spiliopoulos & Hertwig (2019) fit different models to decision-from-experience data from past experiments. Their results suggest that skewness is a predictor of participants' decisions. However, a direct assessment of the role of higher order risk attitudes in the decisions-from-experience versus the decision-from-description mode is not available in the literature.

The current paper aims to provide the first evidence regarding the effects of experience sampling on decision processes for higher order risk attitudes. We combine the separate research streams on decisions from experience and on higher order risk preference (in the description paradigm). Our results support the documented description-experience gap for risk aversion. More importantly, we find a so far undocumented gap for prudence: While prudence is strong in the decisions from description paradigm (as reported in many studies by now, e.g. (Deck &

[1] The fourfold pattern of risk attitudes describes the following behavior: risk aversion for gains with high probability and losses with low probability; risk seeking for gains with low probability and losses with high probability (Tversky & Kahneman, 1992).

[2] A recency effect describes a decision maker's reliance on his or her most recent memories or pieces of information. In sampling experiments, this corresponds to a participant relying on the most recent outcomes s/he sampled.

Schlesinger, 2014; Ebert & Wiesen, 2014; Maier & Rüger, 2011; Noussair et al., 2014), we find no evidence for prudence in decisions from experience.

As Ebert (2013) has shown, the behavioral measures can be related to statistical moment characterizations of higher order risk attitudes. In a second step of our analysis, we make use of these relationships between moments and our behavioral measures by assessing the explanatory power of different statistical moments for the decisions-from-experience data in our data set, as well as in identical analyses for publicly available data for decisions from description. We find that the relative influence of the mean, standard deviation, skewness, and kurtosis on the participants' decisions differs across the two modes, after controlling for sampling error. Our results indicate a weighting shift, with differences in the mean becoming most predictive for decisions, with a more modest role played by skewness and standard deviation, under experience. We explain our finding in terms of the decision processes in decision from experience situations.

1.2 Experimental Setup

We study risk aversion, prudence and temperance in both the description and the experience paradigms. Our setup uses a within-person design, counterbalanced along the description-experience dimension. Moreover, the design includes two between-person conditions, distinguished by the exogeneity, respectively endogeneity, of the number of samples drawn in the decision from experience tasks. In what follows, we present the general experimental setup, then give details on the risky choice tasks, and then describe the experimental procedures and the subject pool.

1.2.1 Design

Participants made a set of binary choices in two within-person parts: A description part, in which they made these decisions from description, and an experience part, in which they made decisions from experience, using the sampling paradigm. Each of the two parts consisted of six independent decision problems. In each problem, participants had to choose between two risky prospects, neutrally framed as "left" and "right" (see Appendix 1.A, figure A1.1 for the presentation of the description task and figures A1.2 and A1.3 for screenshots of the experience part). These six decision problems consisted of two risk aversion tasks, two prudence tasks and two temperance tasks, which are described in detail in the next section. The order of the two parts was counterbalanced: participants' in the description-first order played the description part followed by the experience part, while in the experience-first order the reversed order was used.

In the description part, participants saw a full description of the probabilities and payoffs of both prospects. Hence, in each decision task they had full information concerning the choice at hand. The experience part used the sampling paradigm as described by (Hertwig & Erev, 2009), where participants received no description of the prospects but instead sampled them during a sampling stage that preceded a decision stage. Specifically, in the sampling stage participants sampled the two payoff distributions by clicking on one of two buttons on their screen. Each click on one of these buttons produced a result of the underlying outcome distribution of the chosen prospect. Participants could sample different number of times from the two lotteries if they wished (see details below), in any order. This setup explicitly allows for sampling errors, as there is no guarantee that participants draw a sample that correctly reflects the underlying probability, or that it includes all possible outcomes. We deliberately opted for this design as sampling errors might also occur in natural settings. We are interested how sampling errors impact the decision processes and subsequently risky choices, compared to a description-based, full-information settings. After sampling, participants proceeded to the decision stage in which they had to choose one of the lotteries. At the end of the experiment, one of the 12 decision tasks was randomly selected and played out for real. For example, if a participant chose a risky lottery in the first risk aversion task and this task was determined to be payoff-relevant, the computer drew one of the outcomes from the underlying distribution. The outcome of this task constituted the payoff of the participant.

Our study was divided into two between-subject conditions: exogenous sampling, in which participants had a fixed number of twenty samples per task (distributed over the two prospects according to the participant's sampling preference), and endogenous sampling, in which participants decided themselves how many samples to draw. In the exogenous sampling condition, the sampling phase ended automatically after the twentieth sample and participants entered the decision stage. This stage was highlighted by an explicit warning that the decision in this round was potentially payoff relevant. In the endogenous sampling condition, participants had to sample at least five times, before they could proceed to the decision stage. This was implemented to ensure that participants had at least a rough understanding of the prospects they were about to make a choice between. The minimum of five samples was explicitly announced to the participants. Otherwise, participants could sample up to fifty times and could leave the sampling stage at any time after the fifth sample drawn. The maximum of fifty samples was not explicitly mentioned in the instructions, in order to allow participants to sample freely without setting a reference point for the appropriate number of samples. However, once participants had only five samples left, they saw a warning on each subsequent screen, showing the number of samples they have left. Once the fiftieth sample was drawn, participants proceeded automatically to the decision stage, as in the exogenous sampling condition.[3] The experiment was programmed using z-Tree 3.6.7 (Fischbacher, 2007).

1.2.2 Tasks

Each condition included six different binary choice tasks: two tasks for the elicitation of risk aversion, two for the elicitation of prudence and two for the elicitation of temperance. As

[3] Out of 282 participants, the following actually reached the fiftieth sample in the experience part: 19 in risk aversion task 1, 38 in risk aversion task 2, 27 in prudence task 1, 24 in prudence task 2, 25 in temperance task 1, and 29 in temperance task 2 (see Tasks).

participants played both the description and experience part, each of the 12 prospects in the second part they played was slightly modified by adding 50 Cents to each outcome of these prospects, to obtain two similar but slightly different sets of lotteries. This transformation ensured that the distributions of both options within each task was not repetitive yet kept the same underlying moments. The participants in the description-first condition encountered Set 1 (unmodified lotteries) in the description part, and Set 2 (modified lotteries) in the experience-part. The participants in the experience-first condition encountered Set 1 (unmodified lotteries) in the experience part, and Set 2 (modified lotteries) in the description task. Table 1.1 presents a full description of all prospects.

Each of the risk aversion tasks consisted of a two-outcome risky lottery and a safe payoff. In risk aversion task 1, the high outcome of the risky prospect was associated with a high probability of 90%, while in risk aversion task 2 the high outcome was realized only with a low probability of 10%. The safe payoff was adjusted accordingly to be equal to the expected payoff of the respective lottery. Hence, the lottery and the safe payoff differed only in their riskiness and the variability of the outcomes. The higher-order tasks were constructed according to the framework of (Eeckhoudt & Schlesinger, 2006). For the prudence tasks, an additional mean-zero risk was added to either the low outcome (imprudent option) or the high outcome (prudent option) of a two-outcome prospect. Additionally, we increased the upside risk of the prudent option in prudence task 2, which gave the prudent option an overall higher expected value and skewness. For the temperance tasks, two zero-mean risks were either added to the same outcome (intemperate option) or evenly distributed over the two possible outcomes (temperate option) of a two-outcome prospect. Importantly, all higher-order lotteries were reduced using reduction of compound lotteries (as shown in Table 1.1), since risk apportionment in the setup of (Eeckhoudt & Schlesinger, 2006) is not possible in sampling paradigms. This implies losing the sequential nature inherent in risk apportionment (see Trautmann & van de Kuilen, 2018, for a discussion of the reduced versus the risk-apportionment forms). To ensure comparability of the description and the experience conditions, we therefore used the same reduced format in the description task as well, to allow for a clean comparison between the two conditions.

1.2.3 Laboratory Procedures and Participants

Procedure. In the laboratory, participants were seated in individual cubicles equipped with computers, the general instructions (see online supplement for the full instructions), a receipt for their payment for later use and a pen. We also provided them with an additional sheet, explaining the prospect format used (see online supplement). This sheet was included to make sure that participants understood how prospect distributions work. The sheet also contained a comprehension question. Participants were given three outcome distributions and were asked to sketch the appropriate distribution graph themselves. Their answer was checked by one of the experimenters. If their sketch was correct, they received the instructions for the first part of the experiment and could proceed. Otherwise, they received additional explanation and were asked to try again before proceeding.

After finishing the first part of the experiment, participants received the instructions for the second part and had to confirm again that they have read them before proceeding. Once the participants had finished the second part, they were informed which part and task was chosen for them to be payoff relevant. They also saw the randomly determined outcome of their choice in this task. They answered a short demographic questionnaire (asking for age, gender and field of study) and received their payoff in private.

Participants. Both conditions were run at the AWI lab in Germany. Participants were recruited from a pool of volunteers, mostly consisting of Heidelberg University students, using Hroot (Bock et al., 2014). Participants received a show-up fee of €3 and faced substantial and steep incentives with lottery payoffs ranging from €0 to €20. In the exogenous sampling condition, a total of 182 participants took part, 92 in description-first and 90 in experience-first order. The average age of the participants was 24 years (SD = 4.6), and 53% of them were female. Participants in the exogenous sampling condition earned €11.54 on average.

The endogenous sampling condition included 282 participants, with 175 in the experience-first part and 107 in the description-first part.[4] The average age of participants was 23 years (SD = 4.2), with 54% female and 32% having a background in economics. On average, participants earned €10.71. We deliberately included a higher number of participants in the endogenous sampling condition in order to account for the higher variation in samples drawn in the endogenous sampling process.

1.3 Hypotheses

Numerous studies have shown that the fourfold pattern for risk preferences typically observed in decision from description is reversed for decision from experience (Hertwig et al., 2004). For example, the often-found risk seeking behavior for small probability gains reverses to risk averse behavior in a sampling paradigm. Focusing only on the gain domain in our experiment, we expect to find the same pattern in our risk aversion tasks.

Hypothesis 1: *(Replication) For the high probability gain risky task, a higher proportion of participants chooses the risk averse option in the description part compared to the experience part. For the low probability gain risky task, a lower proportion of participants chooses the risk averse option in the description part compared to the experience part.*

For decisions involving prudent lotteries we expect a similar gap. Given the skewness properties of the two lotteries in these decision problems, the prudence tasks are similar in structure to the low probability gains. On the other hand, the increased complexity of the prudence task lotteries and the larger number of outcomes may induce subjects to sample more extensively in the endogenous sampling condition. However, as the description-experience gap for risk has been shown to be persistent even when the number of samples was increased (Hau et al., 2008;

[4] The discrepancy between the numbers of participants in both treatments was introduced when we conducted more sessions of the experience first treatment due to scheduling issues.

Ungemach et al., 2009) or when participants received stronger incentives to sample (Hau et al., 2008), we expect the same to hold for the prudence tasks. Moreover, if decision makers rely on small and possibly recent samples (Wulff et al., 2018), the structure of the lotteries induces under-sampling of the attractive right branch of the prudent lottery and of the unattractive left branch of the imprudent lottery. This effect would similarly lead to prudent lotteries to initially appearing less attractive to decision makers compared to the description case, inducing underweighting of the upside risk. Prudence task 2 tests this prediction by making the potentially underweighted (in decision from experience) upside risk of the prudent option more attractive. If skew is neglected in sampling, the change should have little effect on valuations.

Hypothesis 2: *In both prudence tasks, a higher proportion of participants chooses the prudent option in the description part compared to the experience part.*

Evidence on temperance in decisions from description is more mixed (Trautmann & van de Kuilen, 2018). For our current interest in description-experience differences, we observe that the lotteries in the temperance decision tasks are symmetric. Asymmetry seems an important ingredient in the description-experience gap, suggesting that differences may be less pronounced for temperance. On the other hand, the fatter tails for intemperance allow decision makers to observe more extreme outcomes, both good and bad ones, in small samples. With bad outcomes framed as losses with respect to average samples and loss aversion, we predict that the temperate option becomes more attractive under experience. This would be true for both endogenous and exogenous sampling.

Hypothesis 3: *In both temperance tasks, a lower proportion of participants chooses the temperate option in the description part compared to the experience part.*

In addition, we are interested in potential differences in the sampling behavior of participants in the two different conditions. First, previous studies endogenizing the number of samples found that participants only drew small samples, around and mostly below 20 (Ert & Trautmann, 2014; Hau et al., 2010; Wulff et al., 2018). We expect to find similar numbers for the risk aversion tasks. Since the higher-order tasks are more complex in regard of the possible number of outcomes in both lotteries, we expect participants to take this into consideration when deciding to stop sampling and proceed to the decision round. If the focus on recent small samples is relevant, this would not affect preferences though.

Hypothesis 4: *Participants draw more samples in the higher-order risk tasks compared to the risk aversion tasks.*

1.4 Results

1.4.1 The Description-Experience Gap

Table 1.2 shows the pooled proportions of participants choosing the risk averse (safe), prudent, or temperate option in the different tasks in the description and experience conditions. In both conditions, we clearly replicate the description-experience gap for binary risky choices: For the high probability gain versus safe choice, the proportion of risk averse choices is higher for description (91% and 93% in the exogenous and endogenous treatments, respectively) compared to experience (47% and 39%). Conversely, in the low probability gain versus safe choice, the number of risk averse decisions increases when switching from description to experience (from 35% and 47% to 72% and 75%, respectively). Hence, we find a similar reversal for risky choice as previous studies (Hertwig & Erev, 2009): risk seeking for low probability high outcome lotteries in description-based choice reverts to risk aversion in experience-based choice; and risk aversion for high probability modest outcome lotteries reverts to risk seeking.

Table 1.2: Proportion (in % of Participants) of Safe, Prudent and Temperate Choices.

	A: Exogenous sampling			**B: Endogenous sampling**			**ΔGap AB**
	Description[1]	Experience[1]	Difference test[2]	Description[1]	Experience[1]	Difference test[2]	Difference test[3]
Risk aversion task 1 (high prob.)	0.91^{***}	0.47	$z(182) = 9.04$ $p < 0.001$	0.93^{***}	0.39^{**}	$z(282) = 13.44$ $p < 0.001$	$t = -1.79$ $p = 0.074$
Risk aversion task 2 (low prob.)	0.35^{***}	0.72^{***}	$z(182) = -7.04$ $p < 0.001$	0.47	0.75^{***}	$z(282) = -6.73$ $p < 0.001$	$t = -1.55$ $p = 0.122$
Prudence task 1	0.64^{***}	0.52	$z(182) = 2.33$ $p = 0.020$	0.68^{***}	0.46	$z(282) = 5.28$ $p < 0.001$	$t = -1.55$ $p = 0.123$
Prudence task 2 (non ES)	0.82^{***}	0.53	$z(182) = 5.82$ $p < 0.001$	0.82^{***}	0.52	$z(282) = 7.62$ $p < 0.001$	$t = -0.26$ $p = 0.798$
Temperance task 1	0.49	0.41^{*}	$z(182) = 1.47$ $p = 0.140$	0.43^{**}	0.37^{***}	$z(282) = 1.37$ $p = 0.169$	$t = 0.32$ $p = 0.748$
Temperance task 2	0.46	0.45	$z(182) = 0.32$ $p = 0.752$	0.47	0.47	$z(282) = 0.00$ $p = 1.000$	$t = 0.27$ $p = 0.790$

[1] Significance according to a binomial test of the proportion being the result of indifference (H_0: choice proportion = 0.5, *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$)
[2] p-value is from a two-sample proportions test
[3] p-value is from an unpaired t-test

The results of the prudence tasks reveal a clear difference as well: Most participants in both conditions chose the prudent option (64%-82%) from description, which corroborates previous findings (e.g., Trautmann & van de Kuilen, 2018). Yet, when they made their decisions from experience, this clear preference disappeared (46%-53%). There is a higher proportion of prudent choices in prudence task 2 compared to prudence task 1 in the description parts, as expected due to our manipulation of the upside risk branch of the prudent lottery. In line with hypothesis 2, the proportion of prudent choices decreased to around 50% in the experience-based condition for prudence task 2, suggesting that (positive) skewness receives less weight in decisions from experience.

For temperance, the proportion of participants choosing the temperate option in both the description and the experience part, in both the exogenous and the endogenous sampling condition, is close to 50%, with a tendency towards intemperance. Results are consistent with previous findings for description-based decisions in the literature (temperance being weaker than prudence in most studies). In Temperance task 1 and mainly in the endogenous sampling condition, we find a stronger tendency towards intemperance. Overall, the proportions of temperate choices in the description and the experience parts do not significantly differ from each other, however. We find no evidence for a description-experience gap for lotteries that differ only in the fatness of their tails. This suggests that asymmetry is an important aspect of description-experience gaps.

Importantly, the proportions of participants' choosing a specific option and the resulting gaps are strikingly similar in both sampling conditions for all tasks, despite the endogenous size of the sample in the second condition, and the resulting large differences in sample across subjects (see Statistical Moments and Decisions from Experience). To test the gap across conditions, we define the gap at the individual level by the difference in choice between the description and the experience parts. The difference tests for ΔGap AB in Table 1.2 indicate no significant differences between endogenous and exogenous. Moreover, results are robust to using between-subjects comparisons on the basis of the first-part decisions of the two different counterbalanced orders instead of within-subjects analysis.

1.4.2 Statistical Moments and Decisions from Experience

To better understand which factors underlie the participants' choices, we analyze how the observed statistical moments map to the observed decisions. This informs us about the underlying cognitive processes, showing how the observed moments of the prospects map into the participants' choices. To do so, we conduct a panel-probit analysis for the experience parts of both conditions. The dependent variable is the probability of choosing the risk averse, prudent or temperate option in a given task. The key variables of interest are the differences (between the two lotteries in a choice problem) in the four central moments of the respective lottery as observed by the participant in her sample: mean, standard deviation, skewness and kurtosis. More precisely, we calculate for each participant the four above mentioned statistical moments based on their overall individual sample for both lotteries in a given task (see Appendix 1.B for details on the calculation of moments). This means that we treat the sample as-if participants memorized each outcome and weighted them the same. For the safe option in risk aversion tasks 1 and 2, the usual measures for skewness and kurtosis are not defined, making it impossible to calculate a difference. In these cases, the skewness and kurtosis are normalized to zero. This normalization captures that the risky options have either a positive or negative skewness or kurtosis versus the safe one, which might make them differently favorable for the decision makers. The regression analysis also includes different controls. All regressions control for age, gender, the order in which a participant played both parts, having an economics background and the specific task through task fixed effects (Controls). While these controls are exogenous with respect to the decision-making process, sampling behavior is endogenous. We control for sampling features by including the number of samples drawn (if endogenous sampling), the number of samples drawn from the right lottery (hence the safe, prudent or temperate option), and whether the participant saw all outcomes of both lotteries (Fox & Hadar, 2006). We call this set of variables Sampling Controls. In the endogenous sampling regressions, we also run a specification that includes the interaction of the number of samples drawn and the differences in the statistical moments, to detect whether differences in sampling affect the correlation between the moment differences and choice behavior. Standard errors are clustered at the subject level.

Table 1.3 shows the estimation results of these models. First, notice that the coefficient sizes and directions are similar in both conditions. There is a strong association of choices with the subjectively sampled mean differences, both in terms of the regressors in the probit model, as well as the marginal effects. The association with the difference in standard deviation is negative and significant, except in model III for the endogenous sampling condition, where we find no significant main effect for standard deviation. The marginal effect sizes for the difference in standard deviation are substantially smaller than those for the difference in the sampled means (Wald test, $\chi^2_1 >= 5.33, p <= 0.021$ in all models). In accordance with Spiliopoulos & Hertwig (2019) results, the association of choices with difference in sampled skewness is positive and significant in all models of both conditions. It has smaller marginal effects compared to the sampled mean difference in most conditions (Wald test, $\chi^2_1 >= 11.85, p < 0.01$ Wald test, $\chi^2_1 = 0.44, p = 0.51$ for Model III), but larger effects than standard deviation difference in the endogenous sampling condition (Wald test, $\chi^2_1 >= 10.74, p < 0.01$; no difference in the exogenous condition $\chi^2_1 <= 2.65, p >= 0.1$). Skewness seeking thus matters for experience-based decision, and apparently more so than the standard variation, but less so than expected payoffs. The regressors for the difference in kurtosis are insignificant and close to zero. Considering the interaction effects in model III, we find significant effects for standard deviation and the mean: a larger number of samples amplifies the main effects. However, all interactions effects exhibit small effect sizes. This indicates that the observed differences are not driven by the interaction of larger samples and higher moment differences.

What influence would such a probit model predict for the various moment differences in decisions from description? We cannot use our own description data, as all participants faced the same tasks and hence there would be no variation. To assess how the statistical moments affect decisions in the description paradigm, we run similar probit regressions on decisions-from-description data of Wulff et al. (2018). We find the same basic pattern as in Table 1.3: Significant positive coefficients for the mean and skewness, a significant negative coefficient for the standard deviation, and a non-significant and close to zero coefficient for the kurtosis (details in Appendix 1.C). As before, the mean and skewness differences have a larger marginal effect than the standard deviation (Wald test, $\chi^2_1 >= 32.95, p < 0.01$). Importantly, the coefficient for skewness differences is significantly larger under the description paradigm than the one for the mean difference (Wald test, $\chi^2_1 = 26.67, p < 0.01$). While Wulff et al. (2018) did not

Table 1.3: Explaining Choices in Decisions from Experience.

	Exogenous Sampling		Endogenous Sampling		
	I	II	I	II	III
Sampled distribution moments					
ΔMean	0.489**	0.473**	0.420**	0.389**	0.179**
	(0.079) [0.144]	(0.079) [0.135]	(0.043) [0.117]	(0.043) [0.101]	(0.063) [0.127]
ΔSD	-0.096**	-0.157**	-0.068**	-0.078**	0.001
	(0.030) [−0.028]	(0.035) [−0.045]	(0.020) [−0.019]	(0.026) [−0.020]	(0.037) [−0.027]
ΔSkewness	0.165**	0.152*	0.247**	0.221**	0.239**
	(0.037) [0.048]	(0.036) [0.043]	(0.033) [0.069]	(0.034) [0.057]	(0.067) [0.065]
ΔKurtosis	-0.004	-0.021	0.008	-0.012	-0.011
	(0.015) [−0.001]	(0.016) [−0.006]	(0.010) [0.002]	(0.011) [−0.003]	(0.033) [−0.003]
Interaction between number of samples and moment differences					
ΔMean*Samples					0.013**
					(0.003)
ΔSD*Samples					-0.04**
					(0.001)
ΔSkewness*Samples					0.001
					(0.002)
ΔKurtosis*Samples					-0.000
					(0.001)
Sampling Controls[1]	No	Yes	Mo	Yes	Yes
Controls[2]	Yes	Yes	Yes	Yes	Yes
N	182	182	282	282	282

Notes: Probit regressions, clustered standard errors in parentheses, marginal effects in brackets (signs of the marginal effects are corroborated by unreported OLS regression analysis). Differences in mean, SD, skewness and kurtosis are defined as the subjectively sampled difference in the respective statistical moment between the right and the left lottery of a task. 1: Sampling controls include a dummy for whether or not the participant saw all possible outcomes, the number of samples from the right lottery and, for endogenous sampling, the total number of samples drawn. 2: Controls include the constant, age, gender, the order in which the parts were played, being an economist and a dummy for each task. *, **, *** indicate significance at the .05,.01 and .001 significance level.

have available prudent choices, the strong role of skewness explains the empirical evidence on the prevalence of prudence (all studies were based on decision from description). Observing the same pattern of moment effects (positive effects of mean and skew, negative effect of variance) on decisions from description and decisions from experience suggest that similar cognitive processes are at play in both decision modes; however, our results show that in decisions from experience, decision weight is shifted from the observed skewness to the observed mean, with substantial effects on observed choices.

1.4.3 Recency

Previous studies explored the concept of recency as a potential contributor to the emergence of a description-experience gap. A participant exhibiting recency relies more on the more recent samples to make a decision (Hogarth & Einhorn, 1992). Results on recency have in general been mixed (Hau et al., 2008; Hertwig et al., 2004; Rakow et al., 2008; Ungemach et al., 2009; Wulff et al., 2018; Wulff & Pachur, 2016). Wulff et al. (2018) highlight the importance of *optional stopping* (Fried & Peterson, 1969; Wald, 1947), i.e. stopping an endogenous sampling process after observing e.g. a rare outcome or a specific stream of outcomes. Their meta-analysis found a persistent recency effect for endogenous sampling conditions, but no consistent recency effect for exogenous sampling conditions. As participants in their data rely on small samples, they interpret this as recency occurring due to *optional stopping*, rather than due to memory limitations (Ashby & Rakow, 2014) or positional weighting (Hogarth & Einhorn, 1992). To test for recency in our data, we use three definitions for recency, as employed by Wulff et al. (2018). According to these three definitions, the samples of the individual participants are split into a primacy and a recency set. This is done by either i) splitting the complete sample into two halves (across), ii) splitting the sample in half for each of the two sampled lotteries (within), or iii) splitting the sample in two halves at the second switching point (mirror image), i.e. the first time an individual returns to a previously sampled lottery. We then calculate the sampled mean for each prospect in both the primacy and the recency sets. As discussed above, the mean is highly predictive of choices in decisions from experience. We then compare whether the sampled mean in the primacy or recency set is the better predictor of the observed choices by counting the number of correct predictions, i.e. in which instances choosing the option with the higher mean would yield the same decision as the actual choice (see Appendix 1.D for the corresponding figures). The results from this analysis indicate that both the recency and the primacy sets performed equally well in predicting actual choices when using the across and mirror image split definitions, in both the exogenous and the endogenous conditions. The within split definition, on the other hand, produced consistent recency effects in both conditions (see Appendix 1.D for further details). Thus, the current data suggest some evidence for recency. This differs from the results of the Wulff et al. (2018) meta-analysis, which only found recency for endogenous sampling conditions. As the number of samples drawn were on average close in both our conditions (see section below), this could potentially explain our finding.

1.4.4 Sampling Behavior

Participants in the endogenous sampling condition chose themselves when to terminate sampling. Figure 1.1 shows the distribution of the samples drawn by the participants for each task, the solid lines mark the respective median amounts of samples. For the risk tasks, the median number of samples were 15 and 19, respectively. Participants drew on average larger samples in the higher-order tasks, with medians ranging from 23-25 ($p <= 0.001$, one-sided matched-pair sign test). That is, participants did not come close to exhausting the full number of possible samples. While the higher order risk tasks exhibit a rather even distribution of samples, the distributions of the risk aversion tasks are bimodal with peaks at the upper and the lower end. This might be caused by the comparison of a sure payoff with a heavily tailed lottery in the risk aversion tasks. Generally, sampling behavior can be influenced by cognitive abilities (memory, rational thinking, numeracy, etc.) or by ecological factors of the decision environment. Examples for the latter used in literature include the presence of losses, the order in which problems are presented and the variability of outcomes (Lejarraga et al., 2012; Wulff et al., 2018). The presence of tail events, such as in our case, could also be an important ecological factor. Table 1.4 shows a test of the different possible ecological moderators on the sample size in our experiment. Only the order in which the two parts were played appears to significantly moderate the number of samples drawn.

Figure 1.1: Distribution of samples drawn (black lines indicate the median).

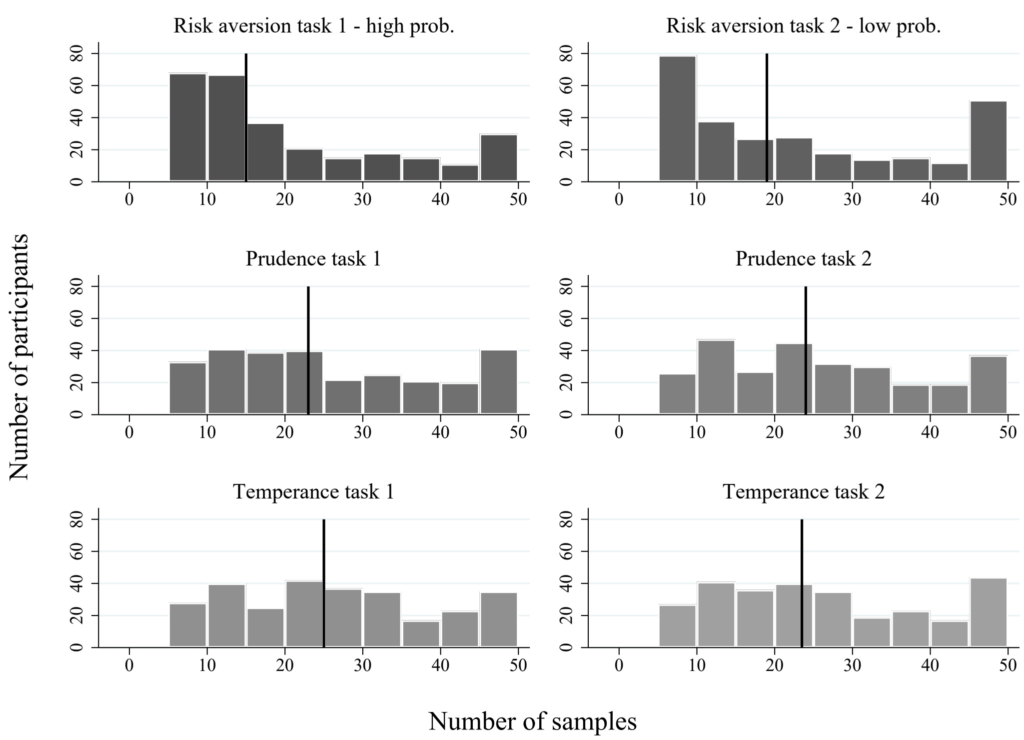


Table 1.5 shows the description-experience gap split along the median number of samples

Table 1.4: Moderators of sample size.

	Number of samples drawn
Avg. Mean	-1.310
	(0.908)
Avg. SD	1.326
	(1.197)
Avg. Skewness	4.684
	(2.939)
Avg. Kurtosis	1.813
	(1.534)
Experience tasks first	-3.249**
	(1.333)
N	282

Notes: Panel regression, clustered standard errors in parentheses. Differences in mean, SD, skewness and kurtosis are defined as the average of the true underlying statistical moments between the right and the left lottery of a task. *, **, *** indicate significance at the .05,.01 and .001 significance level.

drawn in the endogenous sampling condition. For both groups the proportion of risk averse, prudent or temperate decisions are close to each other in the description part but vary in the experience part. For risk aversion, the gap is significantly larger for below median samplers. The effect points in the same direction for prudence, but insignificantly so. For temperance, the effect points in different direction in the two tasks (see column ΔGap AB in Table 1.5).

Potentially, the observed gap may depend on whether participants saw all possible outcomes in each given task (Fox & Hadar, 2006). While a larger sample in general increases the likelihood of a participant observing all possible outcomes, there is still no guarantee that all outcomes are observed. Tables 1.6 and 1.7 therefore split the observations into two subsamples: one in which the participants did not see all possible outcomes of a task, and one in which they did. As the data from the exogenous condition can be split in the same way, these observations are included here. Table 1.6 shows the results for the exogenous sampling condition, while Table 1.7 shows the results for the endogenous sampling condition. The number of observations in each sample show that in many cases only about half of all participants did actually see all possible outcomes.

For the exogenous sampling condition in Table 1.6, no systematic effects emerge. The endogenous sampling condition in Table 1.7 exhibits a similar (but more pronounced) pattern as for splitting observations along the median of drawn samples presented in Table 1.5. This is not surprising as a larger sample increases the likelihood of observing all possible results (Pearson correlation ranging from 0.54 to 0.63 in the different tasks). Although the description-experience gap is significantly emerging for both groups, it is less pronounced for participants who did see

Table 1.5: Difference Between Below and Above Median Samplers.

	A: Below median samplers			B: Above median samplers			ΔGap AB
	Description[1]	Experience[1]	Difference test[2]	Description[1]	Experience[1]	Difference test[2]	Difference test[3]
Risk aversion task 1 (high prob.)	0.93^{***}	0.21^{***}	$z(135) = 12.05$ $p < 0.001$	0.92^{***}	0.57	$z(136) = 6.55$ $p < 0.001$	$t(269) = 5.75$ $p < 0.001$
Risk aversion task 2 (low prob.)	0.49	0.86^{***}	$z(138) = -6.53$ $p < 0.001$	0.45	0.65^{***}	$z(139) = -3.26$ $p < 0.001$	$t(275) = -2.42$ $p = 0.016$
Prudence task 1	0.67^{***}	0.43	$z(135) = 4.04$ $p < 0.001$	0.71^{***}	0.49	$z(138) = 3.69$ $p < 0.001$	$t(271) = 0.34$ $p = 0.733$
Prudence task 2 (non ES)	0.82^{***}	0.51	$z(131) = 5.38$ $p < 0.001$	0.81^{***}	0.55	$z(139) = 4.76$ $p < 0.001$	$t(268) = 0.60$ $p = 0.548$
Temperance task 1	0.44	0.41^{*}	$z(138) = 0.49$ $p = 0.626$	0.42	0.32^{***}	$z(134) = 1.65$ $p = 0.090$	$t(270) = -0.86$ $p = 0.392$
Temperance task 2	0.54	0.44	$z(132) = 1.60$ $p = 0.109$	0.41^{*}	0.51	$z(136) = -1.70$ $p = 0.090$	$t(267) = 2.56$ $p = 0.011$

[1] Significance according to a binomial test of the proportion being the result of indifference (H_0: choice proportion = 0.5, *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$)
[2] p-value is from a two-sample proportions test
[3] p-value is from an unpaired t-test

all possible outcomes compared to those who did not, for both risk and the prudence tasks (see column ΔGap AB). Hence, sampling error at least partly drives the observed results (Fox & Hadar, 2006).

The findings in Tables 1.6 and 1.7 also explain why there are no differences in the description-experience gaps between the exogenous and endogenous sampling conditions: Participants with a very small sample in the experience task (which likely did not see all outcomes) appear to move further away from their choice in description on average. Conversely, those with a larger and more complete sample seem to be much closer to their choice in description. The average choice pattern then remains close to that in the exogenous sampling condition. This explanation is supported by the stochastic properties of the prospects used. A simulation of hypothetical players sampling from the lotteries (see Appendix 1.E) shows that a very small sample leads to underestimation of the expected value of the safe option in risk aversion task 1, the risky option in risk aversion task 2, and the prudent option in prudence task 1 and 2. Given the large role of expected value for choices in decision from experience (see Table 1.3), this pattern matches the description-experience gaps observed for risk and prudence.

Table 1.6: Exogenous - Difference Between Participants Who Saw All Outcomes and Those Who Did Not.

	A: Did not see all outcomes				B: Saw all outcomes				ΔGap AB
	N	Description[1]	Experience[1]	Difference test[2]	N	Description[1]	Experience[1]	Difference test[2]	Difference test[3]
Risk aversion task 1 (high prob.)	54	0.91^{***}	0.47	$z(54) = 6.33$ $p < 0.001$	128	0.90^{***}	0.46	$z(128) = 6.46$ $p < 0.001$	$t(180) = -0.03$ $p = 0.974$
Risk aversion task 2 (low prob.)	94	0.42	0.69^{***}	$z(94) = -3.80$ $p < 0.001$	88	0.28^{***}	0.75^{***}	$z(88) = -6.22$ $p < 0.001$	$t(180) = 2.10$ $p = 0.037$
Prudence task 1	84	0.68^{**}	0.55	$z(84) = 1.81$ $p = 0.071$	98	0.60	0.48	$z(98) = 1.50$ $p = 0.133$	$t(180) = 0.16$ $p = 0.870$
Prudence task 2 (non ES)	97	0.83^{***}	0.55	$z(97) = 4.11$ $p < 0.001$	85	0.81^{***}	0.51	$z(85) = 4.13$ $p < 0.001$	$t(180) = -0.19$ $p = 0.846$
Temperance task 1	85	0.50	0.43^{*}	$z(85) = 0.93$ $p = 0.353$	97	0.48	0.40	$z(97) = 1.15$ $p = 0.250$	$t(180) = -0.1$ $p = 0.919$
Temperance task 2	84	0.49	0.39^{*}	$z(84) = 1.48$ $p = 0.140$	98	0.43	0.51	$z(98) = -1.05$ $p = 0.293$	$t(180) = 1.91$ $p = 0.057$

[1] Significance according to a binomial test of the proportion being the result of indifference (H_0: choice proportion = 0.5, *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$)
[2] p-value is from a two-sample proportions test
[3] p-value is from an unpaired t-test

Table 1.7: Endogenous - Difference between participants who saw all outcomes and those who did not.

	A: Did not see all outcomes				B: Saw all outcomes				ΔGap AB
	N	Description[1]	Experience[1]	Difference test[2]	N	Description[1]	Experience[1]	Difference test[2]	Difference test[3]
Risk aversion task 1 (high prob.)	143	0.94^{***}	0.13^{***}	$z(143) = 13.87$ $p < 0.001$	139	0.91^{***}	0.67^{***}	$z(139) = 5.02$ $p < 0.001$	$t(280) = 9.70$ $p < 0.001$
Risk aversion task 2 (low prob.)	138	0.46	0.94^{***}	$z(138) = -8.69$ $p < 0.001$	144	0.48	0.56	$z(144) = -1.42$ $p = 0.157$	$t(280) = -5.47$ $p < 0.001$
Prudence task 1	133	0.67^{***}	0.34^{***}	$z(133) = 5.40$ $p = 0.071$	149	0.70^{***}	0.58	$z(149) = 2.17$ $p = 0.030$	$t(280) = 2.68$ $p = 0.008$
Prudence task 2 (non ES)	135	0.83^{***}	0.36^{**}	$z(135) = 7.81$ $p < 0.001$	147	0.82^{***}	0.67^{***}	$z(147) = 2.93$ $p = 0.003$	$t(280) = 4.28$ $p < 0.001$
Temperance task 1	118	0.42	0.42^{*}	$z(118) = 0.00$ $p = 1.000$	164	0.43	0.34^{***}	$z(164) = 1.82$ $p = 0.069$	$t(280) = -1.23$ $p = 0.219$
Temperance task 2	143	0.50	0.50	$z(143) = 0.00$ $p = 1.000$	129	0.43	0.43	$z(139) = 0.05$ $p = 1.000$	$t(280) = 0.00$ $p = 1.000$

[1] Significance according to a binomial test of the proportion being the result of indifference (H_0: choice proportion = 0.5, *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$)
[2] p-value is from a two-sample proportions test
[3] p-value is from an unpaired t-test

1.5 Discussion

Research on higher-order risk attitudes has exclusively focused on decisions from description, although descriptions are often unavailable outside the lab. Most situations involving risk taking are associated with at best partial knowledge about the underlying distributions, which are then experienced by observing outcomes over time. This study aims to test whether higher-order risk attitudes generalize to such situations of decisions from experience. In contrast to the widely demonstrated prevalence of prudence in decision from description, we find no significant prudence (nor imprudence) in decisions from experience. That is, we find a substantial description-experience gap for prudence. We find no evidence for temperance, and no description-experience gap for temperance.

Our results also replicate the description-experience gap for risk attitudes found in previous studies. Hence, we assume that our experimental design is also well suited for studying the effect of decisions from experience on the higher-order risk attitudes. Skewed risks and higher-order risk attitudes have been shown to play an important role in many economic and financial decisions (Kraus & Litzenberger, 1976; Noussair et al., 2014; Ebert & Hilpert, 2019; Drerup et al., 2023). Especially prudence, as a measure for precaution in saving (Leland, 1968), bargaining (White, 2008), or auction bidding (Esö & White, 2004) has received close attention in economic research. The current results indicate that a careful assessment of the psychological foundations of prudent behavior is needed to fully understand decision processes in the wild. Researchers also need to critically evaluate the environment in which decisions take place. When information is not readily available and has to be acquired by experience, established behavioral patterns might vanish. Precautionary behavior and skewness seeking, as measured by prudence, seem less pronounced in these cases. This could have implications for real-life decision environments where the notion of precaution is particularly prevalent, such as healthcare, drug admission, or individual insurance choice. If prudence is weakened by processing experienced information, this could lead to less precautionary behavior in many important decisions.

We employ higher order statistical moments to explain the participants' experienced-based choices. The results indicate that the subjectively sampled mean payoff is the most relevant predictor of choices between lotteries in the decisions from experience paradigm. Participants seem to have consistently chosen the lottery for which they observed the higher mean. While skewness seeking and aversion to standard deviation also significantly correlate with choices, their marginal effects are substantially smaller than for the effect of the mean; but skewness is more relevant than standard deviation for endogenous sampling. Moments like standard deviation, skewness and kurtosis, which are all measures for the variability of a distribution, should be intuitively easier to experience by sampling. Hence, we would expect a stronger effect. In contrast, analyses of a comparison data set (Wulff et al., 2018) showed that the effect of skewness relative to mean is stronger in decision from description, consistent with the strong prevalence of prudence observed in this paradigm. Overall, our results are consistent with recent studies that also find a strong predictive effect of sampled means (Wulff et al., 2018). Hau et al. (2008) used several models and heuristics to explain results from experience-sampling experiments. The three best performing models were the maximax heuristic (choosing the lottery with the highest possible outcome), prospect theory with parameters fitted to past experience-sampling studies, and the natural mean heuristic (choosing the option with the higher mean). Erev et al. (2010) ran a choice-prediction competition to explain experimental data on experience-sampling choices. The natural mean heuristic once more performed remarkably well. While our participants indeed appear to be strongly influenced by the mean, we find that observed standard deviation and especially skewness also matter. While we find clear effects for the observed moments, we do not know the exact beliefs about the prospects held by the participants (risk perception). If risk perception deviates from the observed moments, sharper predictions will require the elicitation of explicit distributional beliefs. Our results on recency suggest that this may be the case.

Similar to other studies (Ert & Trautmann, 2014; Golan & Ert, 2015; Hau et al., 2008, 2010; Hertwig et al., 2004; Wulff et al., 2018), we find that participants rely on small samples. Participants drew more samples when deciding between two non-degenerate lotteries, compared to deciding between a lottery and a safe option. Hence, participants seem to react to more complex decisions by adjusting the number of samples drawn, even though the overall sample is typically still small. Smaller samples in general carry a risk of missing information by not observing certain outcomes (Hertwig et al., 2006). This introduces a potential sampling error into participants' information sets that can distort their choices (Fox & Hadar, 2006). Indeed, we find that the gaps found in our data tend to be more pronounced for participants who sampled less in the endogenous sampling condition and did not see all outcomes. That is, description-experience gaps seem to derive from a combination of sampling error and the decision to terminate sampling at a specific point in the endogenous sampling condition. Our result of no description-experience gap for the equally complex (in)temperance lotteries suggests that skewness is an important factor for the gap to occur.

There are important open questions. First, how would the presence of rare losses impact people's choices? Losses can moderate the description-experience gap for risky choice (Wulff et al., 2018). Experiencing a rare substantial loss could likewise interact with higher-order risk attitudes. Page et al. (2014) found in a natural field experiment that people suffering major loss during a flood exhibit higher risk seeking behavior afterwards. Notably, the risky lottery they employed to measure risk attitude was highly skewed, which could also indicate more prudent behavior. Potentially, a large loss could highlight the necessity of precautionary measures, such as better insurance coverage (Browne & Hoyt, 2000; Jiang et al., 2013). Second, which criteria cause people to terminate sampling? Even in the endogenous sampling condition, there existed a minimum and a maximum number of samples. In decision situations outside the laboratory, there is typically little guidance as to how much information should be collected.

In some practical settings, people receive information in the description format, but then experience outcomes by observing samples. This is typically the case with investment decisions, and it may lead to decision that look unsatisfactory ex-post after experiencing outcomes, because the initial decision does not reflect the nature of the experience sampling after an investment is made. One approach to account for this effect could be to shift experience to the stage were the decision is actually made, by letting people sample ex-ante from an already known distribution (Goldstein et al., 2008; Kaufmann et al., 2013). This would mirror situations such as receiving descriptive information of an asset from past data, but then collecting own experiences by holding it. Table A1.2 in Appendix 1.C provides a first look at this approach by analyzing publicly available data collected by (Erev et al., 2017). The patterns obtained for the different moments are similar to our experienced-based results. An important extension would be to embed this approach in our direct choice-based higher-order risk attitude setting, i.e., using the explicit lotteries designed to elicit higher-order risk attitudes, rather than the broad set of lotteries in (Erev et al., 2017). Such approaches seem especially promising in the presence of skewed risks as often found in important investment, insurance, or health decisions.

Chapter 1 References

Ashby, N. J., & Rakow, T. (2014). Forgetting the past: Individual differences in recency in subjective valuations from experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1153. (Publisher: American Psychological Association)

Bock, O., Baetge, I., & Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, *71*, 117–120.

Browne, M. J., & Hoyt, R. E. (2000). The demand for flood insurance: empirical evidence. *Journal of risk and uncertainty*, *20*(3), 291–306. (Publisher: Springer)

Brunnermeier, M. K., Gollier, C., & Parker, J. A. (2007). Optimal beliefs, asset prices, and the preference for skewed returns. *American Economic Review*, *97*(2), 159–165.

Deck, C., & Schlesinger, H. (2014). Consistency of higher order risk preferences. *Econometrica*, *82*(5), 1913–1943.

de Roon, F., & Karehnke, P. (2018). Spanning tests for assets with option-like payoffs: The case of hedge funds. *Available at SSRN 2667473*.

Drerup, T. H., Wibral, M., & Zimpelmann, C. (2023). Skewness expectations and portfolio choice. *Experimental Economics*, *26*(1), 107–144.

Ebert, S. (2013). Moment characterization of higher-order risk preferences. *Theory and Decision*, *74*(2), 267–284.

Ebert, S., & Hilpert, C. (2019). Skewness preference and the popularity of technical analysis. *Journal of Banking & Finance*, *109*, 105675.

Ebert, S., Nocetti, D. C., & Schlesinger, H. (2017). Greater mutual aggravation. *Management Science*, *64*(6), 2809–2811.

Ebert, S., & Wiesen, D. (2014). Joint measurement of risk aversion, prudence, and temperance. *Journal of Risk and Uncertainty*, *48*(3), 231–252.

Eeckhoudt, L., & Gollier, C. (2005). The impact of prudence on optimal prevention. *Economic Theory*, *26*(4), 989–994.

Eeckhoudt, L., & Kimball, M. (1992). Background risk, prudence, and the demand for insurance. In *Contributions to insurance economics* (pp. 239–254). Springer.

Eeckhoudt, L., & Schlesinger, H. (2006). Putting risk in its proper place. *The American Economic Review*, *96*(1), 280–289.

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological review*, *124*(4), 369. (Publisher: American Psychological Association)

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

Ert, E., & Trautmann, S. T. (2014). Sampling experience reverses preferences for ambiguity. *Journal of Risk and Uncertainty*, *49*, 31–42.

Esö, P., & White, L. (2004). Precautionary bidding in auctions. *Econometrica*, *72*(1), 77–92.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, *10*(2), 171–178.

Fox, C. R., & Hadar, L. (2006). "Decisions from experience"= sampling error+ prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making*, *1*(2), 159–161.

Fried, L. S., & Peterson, C. R. (1969). Information seeking: Optional versus fixed stopping. *Journal of Experimental Psychology*, *80*(3p1), 525.

Golan, H., & Ert, E. (2015). Pricing decisions from experience: The roles of information-acquisition and response modes. *Cognition*, *136*, 9–13.

Goldstein, D. G., Johnson, E. J., & Sharpe, W. F. (2008). Choosing outcomes versus choosing products: Consumer-focused retirement investment advice. *Journal of Consumer Research*, *35*(3), 440–456.

Harvey, C. R., & Siddique, A. (2000). Conditional skewness in asset pricing tests. *The Journal of Finance*, *55*(3), 1263–1295.

Hau, R., Pleskac, T. J., & Hertwig, R. (2010). Decisions from experience and statistical probabilities: Why they trigger different choices than a priori probabilities. *Journal of Behavioral Decision Making*, *23*(1), 48–68.

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, *21*(5), 493–518.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological science*, *15*(8), 534–539.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2006). The role of information sampling in risky choice. *Information sampling and adaptive cognition*, 72–91.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, *13*(12), 517–523.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive psychology*, *24*(1), 1–55. (Publisher: Elsevier)

Jiang, Y. Y., Asai, Y., & Moridaira, S. (2013). On household insurance demand and loss control: evidence from the Great East Japan Earthquake. *International Journal of Business*, *18*(4), 332. (Publisher: Premier Publishing, Inc.)

Kaufmann, C., Weber, M., & Haisley, E. (2013). The role of experience sampling and graphical displays on one's investment risk appetite. *Management science*, *59*(2), 323–340.

Kimball, M. S. (1990). Precautionary Saving in the Small and in the Large. *Econometrica: Journal of the Econometric Society*, 53–73.

Knight, F. H. (1921). Risk, uncertainty and profit. *New York: Hart, Schaffner and Marx*.

Kraus, A., & Litzenberger, R. H. (1976). Skewness preference and the valuation of risk assets. *The Journal of finance*, *31*(4), 1085–1100.

Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition*, *124*(3), 334–342.

Maier, J., & Rüger, M. (2011). Experimental evidence on higher order risk preferences with real monetary losses. *Chapman University (ESI), Working Paper*, 15–10.

Mao, J. C. (1970). Models of capital budgeting, EV vs ES. *Journal of financial and quantitative analysis*, *4*(05), 657–675.

Menezes, C., Geiss, C., & Tressler, J. (1980). Increasing downside risk. *The American Economic Review*, 921–932.

Noussair, C. N., Trautmann, S. T., & Van de Kuilen, G. (2014). Higher order risk attitudes, demographics, and financial decisions. *Review of Economic Studies*, *81*(1), 325–355.

Page, L., Savage, D. A., & Torgler, B. (2014). Variation in risk seeking behaviour following large losses: A natural experiment. *European Economic Review*, *71*, 121–131. (Publisher: Elsevier)

Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes*, *106*(2), 168–179.

Schneider, S., & Sutter, M. (2020). Higher Order Risk Preferences: New Experimental Measures, Determinants and Related Field Behavior. *Working Paper, Max-Planck-Insitiute Bonn*.

Spiliopoulos, L., & Hertwig, R. (2019). Nonlinear decision weights or moment-based preferences? A model competition involving described and experienced skewness. *Cognition*, *183*, 99–123.

Trautmann, S. T., & van de Kuilen, G. (2018). Higher order risk attitudes: A review of experimental evidence. *European Economic Review*, *103*, 108–124.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*, 297–323.

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, *20*(4), 473–479.

Wald, A. (1947). Foundations of a general theory of sequential decision functions. *Econometrica, Journal of the Econometric Society*, 279–313.

White, L. (2008). Prudence in bargaining: The effect of uncertainty on bargaining outcomes. *Games and Economic Behavior*, *62*(1), 211–231.

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological bulletin*, *144*(2), 140.

Wulff, D. U., & Pachur, T. (2016). Modeling valuations from experience: A comment on Ashby and Rakow (2014).

Chapter 1 Appendix

1.A Presentations and Screen Shots

An example of the description task is depicted in Figure A1.1. Figure A1.2 and A1.3 show an example of the sampling procedure. Participants were presented with two buttons "left" and "right". A click on them produced a randomly drawn result from the underlying distribution. Both screenshots were taken from the endogenous sampling condition and therefore included the option to quit sampling (which appeared after sampling five times) and proceed to the final decision.

Figure A1.1: Example lottery – set 1, risk task 1.

LEFT		RIGHT	
Probability	Outcome	Probability	Outcome
90%	15.00€	100%	13.50€
10%	0.00€		

Figure A1.2: Sampling screen endogenous sampling: lottery buttons.

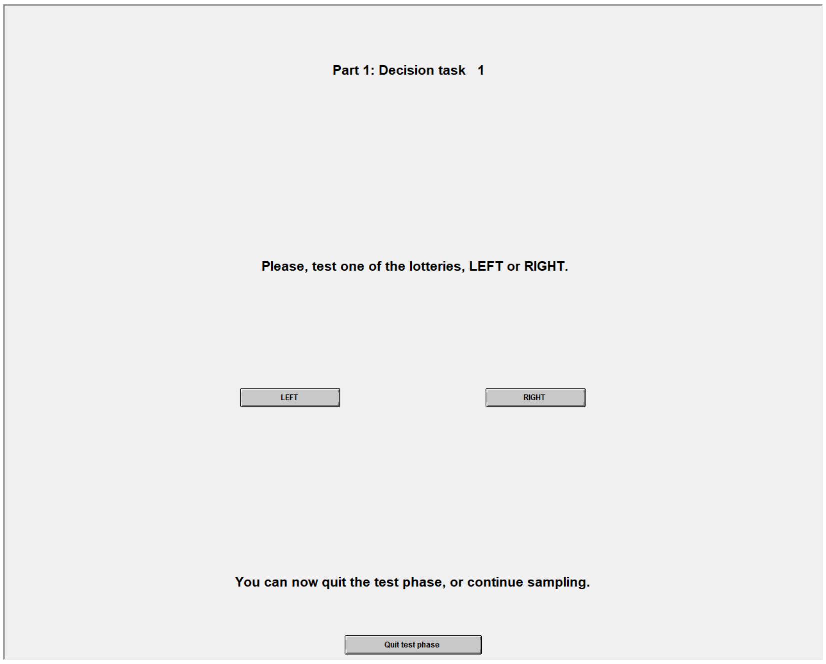
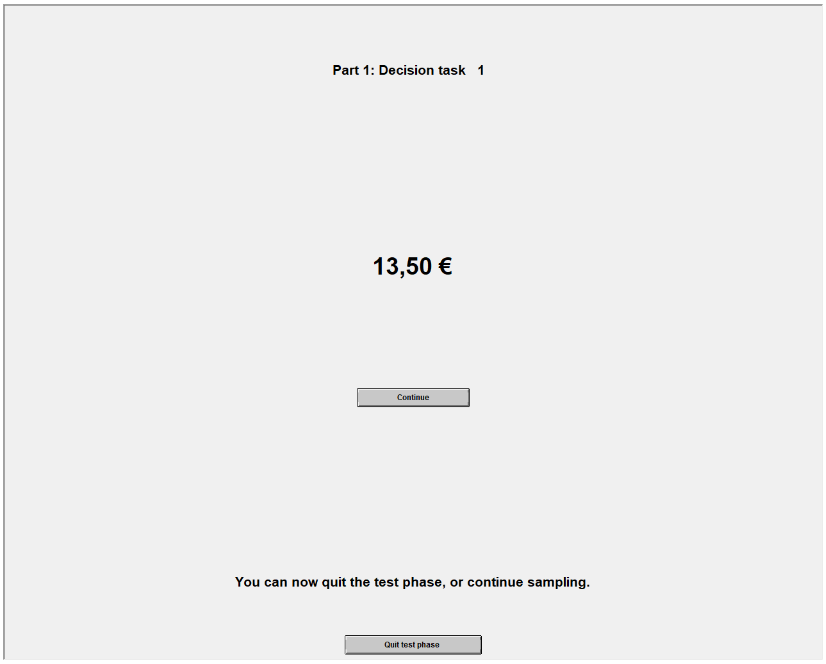


Figure A1.3: Sampling screen endogenous sampling: example outcome.



1.B Calculation of Moments

For the calculation of the subjectively samples moments we used the inbuilt functions of Stata's *egen* command, namely the subcommands mean(), sd(), skew() and kurt(). These subcommands calculate the empirical measures for the mean, the standard deviations, the skewness and the kurtosis, respectively, according to the following formulas:

Mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Standard deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Skewness (third moment):

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

Kurtosis (fourth moment):

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4$$

Note: The standard deviation in both skewness and kurtosis is calculated with the following formula for the standard deviation: $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$. This is because the calculation of skewness and kurtosis is defined over the calculation of distribution moments, with the rth moment of a distribution given as: $m_r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^r$.

1.C Analysis of Moment Effects in Decision from Description Tasks with and without Feedback

Table A1.1 shows results for the role of moments in decision from description tasks. Our analysis uses the data from the meta-analysis of Wulff et al. (2018), available from https://www.dirkwulff.org/. To better make the results comparable, we only use tasks in the gain domain, as our task also did not include losses. This gives us a total of 18,058 observations with 374 subjects/independent observations. We calculate the statistical moments and their differences of the lotteries using the same procedure for Table 3. As the data set provides no demographic information, we cannot control for it. Additionally, the tasks collected in the meta-analysis uses a wider range of outcomes (from 0 to 4876.19), which influences the coefficient size. We thus also provide marginal effects.

Table A1.1: Predicting Choices in Decisions from Description.

	Choosing the safer option
ΔMean	0.006**
	(0.000) [0.002]
ΔSD	−0.006**
	(0.000) [−0.001]
ΔSkewness	0.033**
	(0.005) [0.012]
ΔKurtosis	0.000
	(0.001) [0.000]
N	374

Notes: Probit regression, clustered standard errors in parentheses, marginal effects in brackets. Differences in mean, SD, skewness and kurtosis are defined as the actual differences in the respective statistical moment between the right and the left lottery of a task. *, **, *** indicate significance at the .05,.01 and .001 significance level.

Table A1.2 shows a similar regression for data from the 2015 Choice Prediction Competition (CPC) of Erev et al. (2017). Participants in this data set made decisions for each prospect repeatedly. They knew the description of the two prospects in each task, but also received feedback in the form of a randomly drawn outcome from each option after each choice. Hence, this format can be viewed as sampling from a known distribution, i.e. description. We control for age, gender and the number of the trial and run the regression for the last five trials, after participants could already observe feedback for some time (similar to the regressions in the main text). The coefficients and marginal effects obtained exhibit the same pattern as in Table 3, indicating that even if the description is fully known, sampling influences the participants' weighting of the different risk aspects as captured by the prospect's statistical moments.

Table A1.2: Predicting Choices in Decisions from Description with Feedback.

	Choosing the safer option
ΔMean	0.185**
	(0.008) [0.058]
ΔSD	−0.034**
	(0.002) [−0.011]
ΔSkewness	0.050**
	(0.006) [0.016]
ΔKurtosis	−0.002**
	(0.000) [−0.001]
N	215

Notes: Probit regression, clustered standard errors in parentheses, marginal effects in brackets. Differences in mean, SD, skewness and kurtosis are defined as the actual differences in the respective statistical moment between the right and the left lottery of a task. *, **, *** indicate significance at the .05,.01 and .001 significance level.

1.D Recency and Primacy

The figures below show a graphical representation of the degree of recency vs. primacy in our data, for the three definitions of recency defined in the main text. Figure A1.4 contains the results for the exogenous condition, Figure A1.5 the data for the endogenous condition. Each colored point in the graph for some decision task corresponds to one of the three definitions for splitting the data in a recency and a primacy set. These values are calculated as in Wulff et al. (2018) by subtracting the percentage of choices in line with the recency set from the percentage of choices in line with the primacy set. If both explained the observed choices to an equal degree, this difference should be zero.

Note that we analyze here which option a decision maker only interested in the mean payoff would choose. The graphs A1.4 and A1.5 show the percentage of the actual choices indicated by the recency set minus the percentage indicated by the primacy set. To further break down potential primacy or recency effects, we conducted a panel probit analysis to test to which degree the actual choice is predicted by the different primacy and recency definitions. The results of the analysis are shown in Table A1.3. We find a significant influence of recency according to the within definition in both sampling conditions. The endogenous sampling condition additionally exhibits a significant influence of the primacy across set. To get a better measure how accurately the sets matched the actual choices, we conducted a series of McNemar tests. The McNemar test is a test for homogeneity of paired nominal data. Under the Null hypothesis, the number of Left and Right choices should be the same in the actual choices and as predicted by the different primacy/recency definitions. The p-values of the tests (corrected with the Bonferroni-Holm method), are shown in Tables A1.4 and A1.5. The predictions of the primacy/recency definitions do not match the actual choices quite often. However, for the recency within definition, the McNemar test rejects the null hypothesis only in one instance (endogenous, condition, prudence task 2).

Figure A1.4: Recency in the exogenous sampling condition.

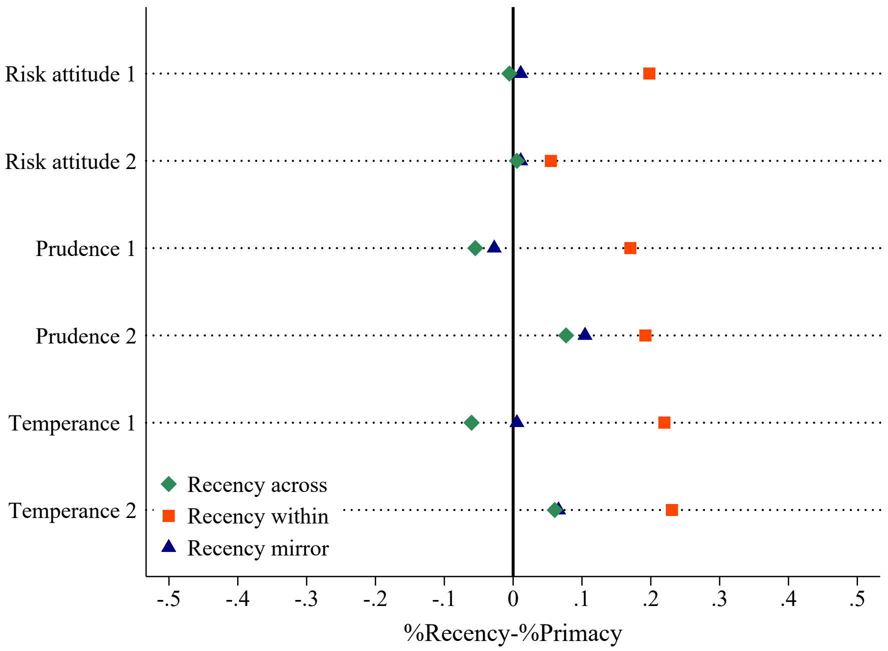


Figure A1.5: Recency in the endogenous sampling condition.

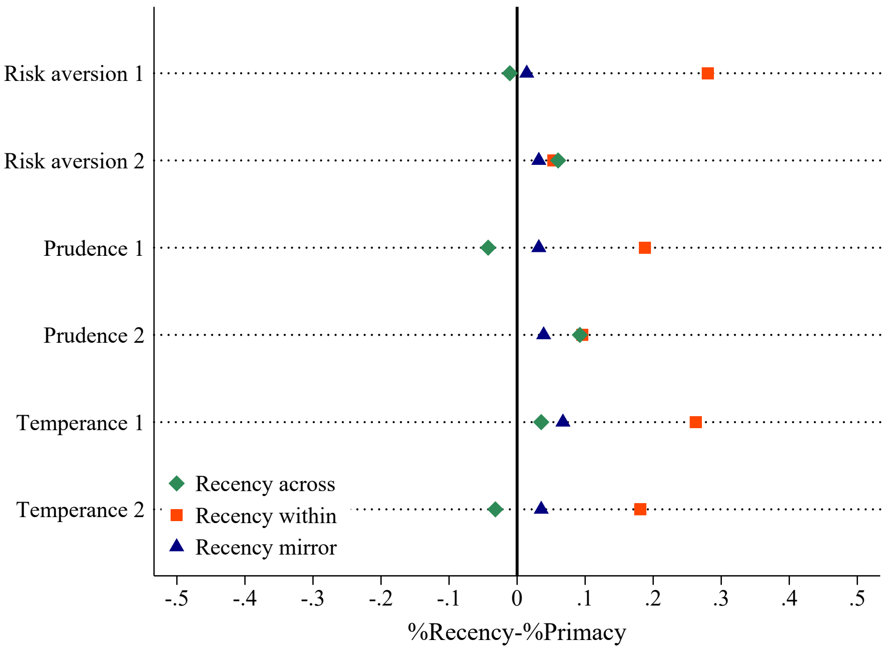


Table A1.3: Predicting Choices in Decisions from Experience with the Primacy and Recency Sets.

	Exogenous	Endogenous
Choices as predicted the primacy/recency sets		
Primacy Across	0.089	0.268**
	(0.137)	(0.085)
Primacy Within	0.048	0.014
	(0.133)	(0.096)
Primacy Mirror	0.113	0.080
	(0.129)	(0.090)
Recency Across	-0.194	0.129
	(-0.139)	(0.093)
Recency Within	1.092**	0.925**
	(0.108)	(0.090)
Recency Mirror	0.139	0.187*
	(0.156)	(0.006)
Sampling Controls[1]	Yes	Yes
Controls[2]	Yes	Yes
N	182	282

Notes: Probit regressions, clustered standard errors in parentheses. 1: Sampling controls include a dummy for whether or not the participant saw all possible outcomes, the number of samples from the right lottery and, for endogenous sampling, the total number of samples drawn. 2: Controls include the constant, age, gender, the order in which the parts were played, being an economist and a dummy for each task. *, **, *** indicate significance at the .05,.01 and .001 significance level.

Table A1.4: McNemar test on differences between prediction and actual choices - exogenous.

	Primacy Across	Primacy Within	Primacy Mirror	Recency Across	Recency Within	Recency Mirror
Risk aversion task 1 (high prob.)	0.415	1	1	1	1	< 0.001
Risk aversion task 2 (low prob.)	0.714	< 0.001	0.015	0.456	1	1
Prudence task 1	0.924	< 0.001	0.461	0.079	1	< 0.001
Prudence task 2 (non ES)	< 0.001	< 0.001	0.001	< 0.001	1	< 0.001
Temperance task 1	< 0.001	< 0.001	< 0.001	< 0.001	1	< 0.001
Temperance task 2	< 0.001	< 0.001	< 0.001	< 0.001	1	< 0.001

Notes: Significance according to a McNemar test of the actual choice and the prediction being similar, p-values corrected by the Bonferroni-Holm method.

Table A1.5: McNemar test on differences between prediction and actual choices - endogenous.

	Primacy Across	Primacy Within	Primacy Mirror	Recency Across	Recency Within	Recency Mirror
Risk aversion task 1 (high prob.)	1	0.006	< 0.001	1	1	0.001
Risk aversion task 2 (low prob.)	1	0.051	1	1	1	1
Prudence task 1	1	< 0.001	1	0.019	1	< 0.001
Prudence task 2 (non ES)	0.563	< 0.001	0.047	0.002	0.015	< 0.001
Temperance task 1	< 0.001	< 0.001	< 0.001	< 0.001	1	< 0.001
Temperance task 2	0.001	< 0.001	< 0.001	0.001	0.180	< 0.001

Notes: Significance according to a McNemar test of the actual choice and the prediction being similar, p-values corrected by the Bonferroni-Holm method.

1.E Simulation

Figures A1.6-A1.11 show the results of simulating 100 hypothetical decision makers drawing samples from our different tasks. Each of these 100 decision makers was simulated to draw $N = 1, 2, 3, ..., 50$ samples from each of the two lotteries offered in each of the tasks. For every simulated set sampled in this way, the observed mean of each option was calculated. The sampled means of both options were then compared with each other. This procedure gives an overview of how a hypothetical sampler drawing equal-sized samples of size N from both options would have perceived the attractiveness of these options in terms of their mean. We ran 100 simulations (each with the aforementioned 100 samplers) and averaged over them. For conciseness we will only show graphs sampled from set 1 of our tasks (see Table 1).

In Figure A1.6, the safe option (dark grey) starts out with having a substantially lower sampled mean on average, compared to the risky option (light grey). This is to be expected: As the risk option offered a very high outcome (€15) with a large probability (90%), the risky option on average dominates the safe option with a payoff of €13.5 before the bad outcome occurs. Once a simulated sampler draws more samples, the likelihood of observing the low outcome (€0 with a 10% chance) increases, which reduces the likelihood of the risky option dominating the safe option. With more samples drawn, the likelihood of one option dominating the other in terms of the mean varies around 50%, with spikes around this value depending on the number of samples due to combinatorics. For risk task 2 the reverse pattern holds (Figure A1.7). For small number of samples, these patterns are consistent with the behavior observed for decisions from experience.

For prudence we observe that, similar to the case of the right-skewed risky lottery, for small samples the prudent lottery (dark grey) performs worse than the imprudent lottery in terms of expected value (figures A1.8 and A1.9). This is true for both prudence decision problems. Again, the pattern is consistent with the behavior observed for decisions from experience. For temperance no such effect is observed. This is consistent with the description-experience gap being closely related to the skewness of the lottery options.

Figure A1.6: Simulation of risk task 1.

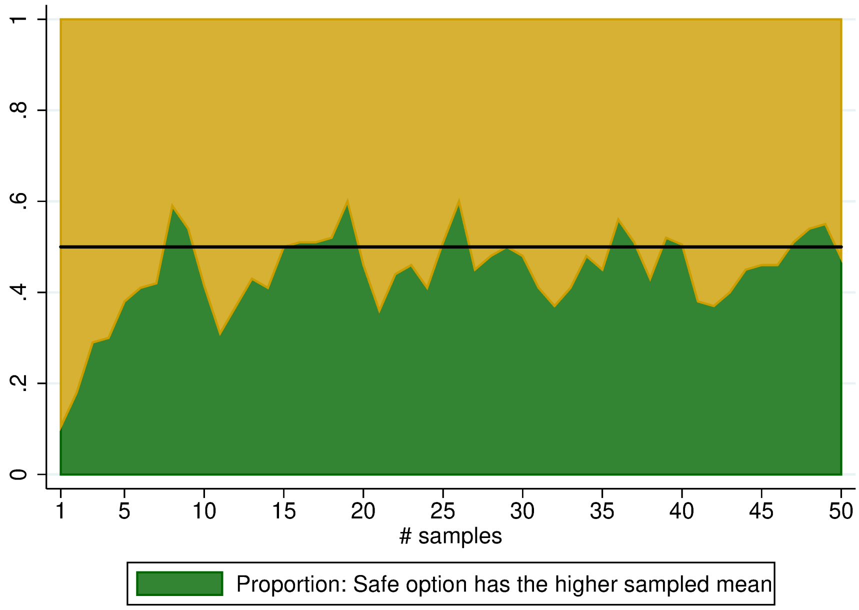


Figure A1.7: Simulation of risk task 2.

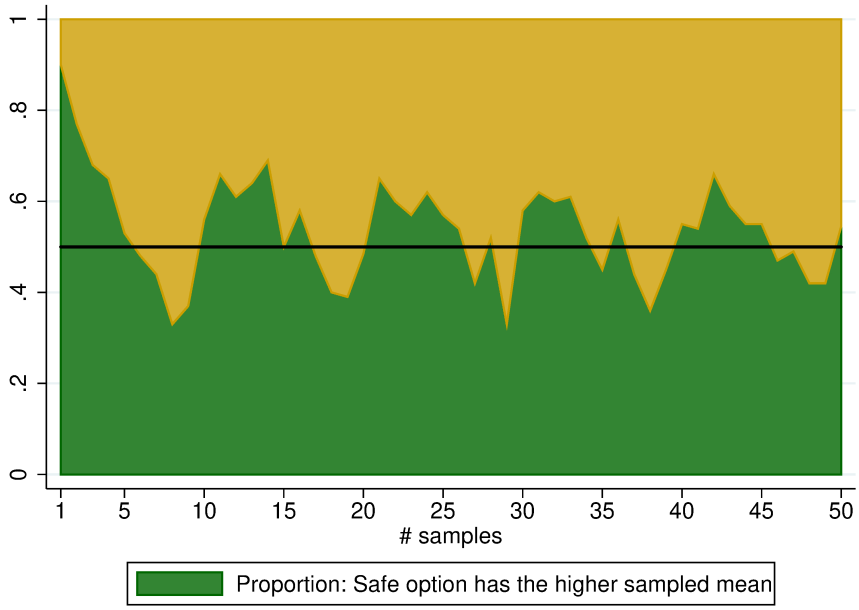


Figure A1.8: Simulation of prudence task 1.

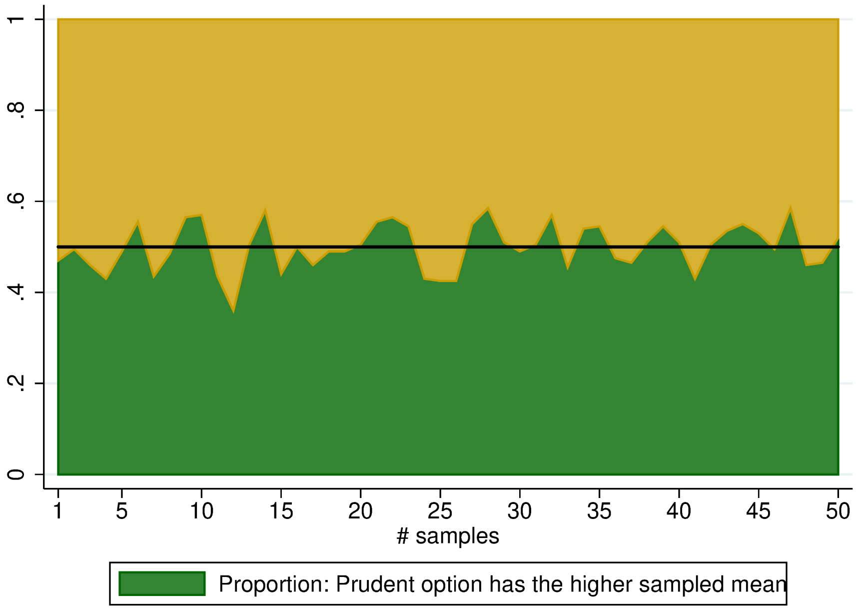


Figure A1.9: Simulation of prudence task 2.

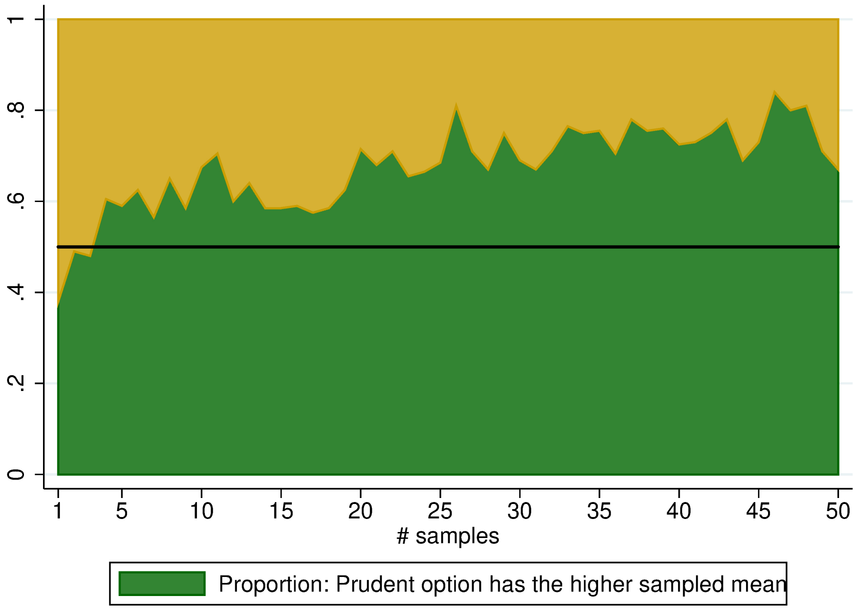


Figure A1.10: Simulation of temperance task 1.

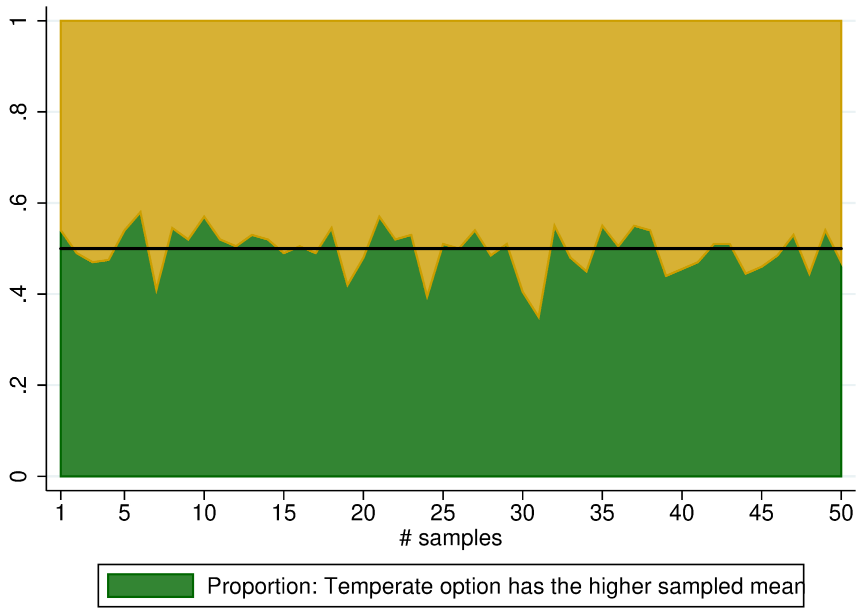
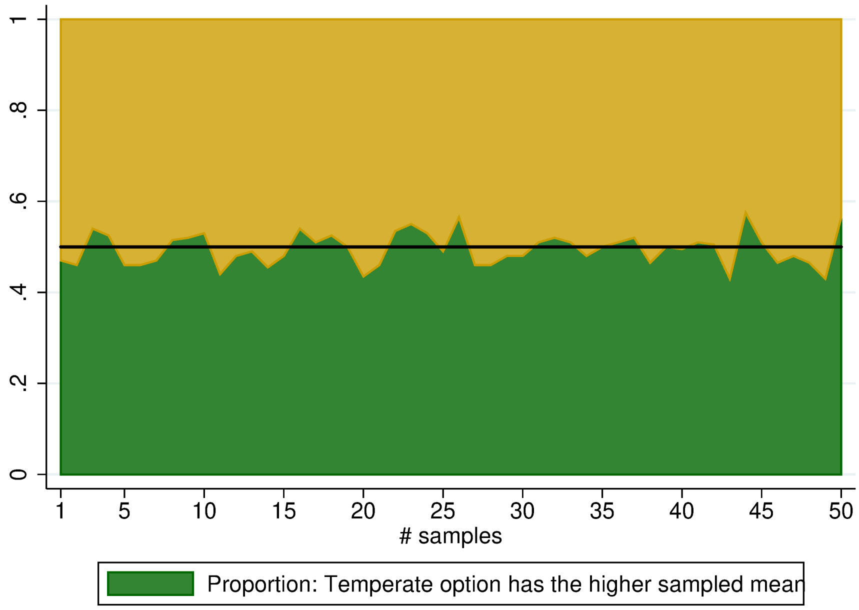


Figure A1.11: Simulation of temperance task 2.



1.F Experimental instructions - German original

Allgemeine Instruktionen

Liebe Studienteilnehmerin,
Lieber Studienteilnehmer,

herzlich Willkommen und vielen Dank für Ihre Teilnahme an unserem heutigen Experiment. Bitte stellen Sie während des Experiments sicher, dass:

- Sie nicht mit anderen Teilnehmern sprechen.
- Ihr Mobiltelefon ausgeschaltet ist.
- Sie keine Notizen machen.

Sollten Sie während des Experiments Fragen haben, heben Sie bitte Ihre Hand. Ein Studienleiter wird dann zu Ihrem Platz kommen, um Ihre Frage zu beantworten.

Dieses Experiment besteht aus zwei Teilen. In jedem Teil werden Sie eine Entscheidung in 6 verschiedenen Situationen, kurz Entscheidungssituationen, treffen. Am Ende des Experiments wird einer der beiden Teile zufällig ausgewählt. Aus diesem Teil wiederum wird eine Entscheidungssituation zufällig ausgewählt und Sie erhalten das Ergebnis Ihrer Entscheidung in dieser Situation ausgezahlt. Zusätzlich erhalten Sie eine sichere Auszahlung von 3€. Ihre gesamte Auszahlung am Ende des Experiments beträgt daher:

Gesamte Auszahlung = Auszahlung der zufällig gewählten Entscheidungssituation des zufällig gewählten Teils + 3€.

Betrachten Sie bevor wir mit dem Experiment beginnen bitte das Informationsblatt zu unsicheren Lotterien und beantworten Sie die Verständnisaufgabe. Haben Sie dies erledigt, heben Sie bitte die Hand, ein Studienleiter wird anschließend Ihre Antwort kurz überprüfen. Danach erhalten Sie die Anleitung für Teil 1 und können mit dem Experiment fortfahren. Sobald Sie Teil 1 abgeschlossen haben, heben Sie erneut die Hand um die Anleitung für Teil 2 zu erhalten. Sie durchlaufen das gesamte Experiment in eigener Geschwindigkeit, unabhängig von den anderen Teilnehmern.

Den gesamten Ablauf des Experiments ist auf der Rückseite dieser Anleitung nochmal kurz dargestellt.

Kurzablauf

In jeder Runde müssen Sie sich zwischen den Lotterien LINKS und RECHTS entscheiden. Eine zufällig gewählte Entscheidungssituation aus den zwei Teilen bestimmt am Ende Ihre Auszahlung.

Allgemeine Instruktionen + Verständnisaufgabe

Heben Sie die Hand, um die Verständnisaufgabe überprüfen zu lassen und die Anleitung für Teil 1 zu erhalten

Teil 1

Heben Sie die Hand, um die Anleitung für Teil 2 zu erhalten.

Teil 2

Ziehung des relevanten Teils und der relevanten Entscheidungssituation

Ergebnisbildschirm

Ausfüllen der Quittung

Fragebogen

Wenn Sie Fragen haben, heben Sie bitte die Hand.

Anleitung Teil 1

Teil 1 des Experiments besteht aus 6 Entscheidungssituationen, in denen Sie sich zwischen den Lotterien LINKS und RECHTS entscheiden. Beide Lotterien führen zu einer Auszahlung, welche jeweils von einer Wahrscheinlichkeitsverteilung abhängt. Ein Beispiel hierfür können Sie unten sehen. Dieses Beispiel wird Ihnen im Experiment nicht als Entscheidungssituation begegnen, sondern dient lediglich der Demonstration.

Beispiel-Entscheidungssituation

LINKS		**RECHTS**	
Wahrscheinlichkeit	Auszahlung	Wahrscheinlichkeit	Auszahlung
50%	**10,00€**	**25%**	**12,50€**
25%	**7,50€**	**25%**	**7,50€**
25%	**2,50€**	**50%**	**5,00€**

Bitte entscheiden Sie sich für LINKS oder RECHTS

Würden Sie sich in diesem Beispiel für LINKS entscheiden, würden Sie mit 50% Wahrscheinlichkeit 10€, mit 25% Wahrscheinlichkeit 7,50€ und mit 25% Wahrscheinlichkeit 2,50€ gewinnen. Würden Sie RECHTS wählen, erhielten Sie mit 25% Wahrscheinlichkeit 12,50€, mit 25% Wahrscheinlichkeit 7,50€, und mit 50% Wahrscheinlichkeit 5€.

Wie zuvor erwähnt kann eine der 6 Entscheidungssituationen aus Teil 1 für Ihre Auszahlung relevant sein. Nachdem Sie alle 6 Entscheidungen getroffen haben wird der Computer zufällig eine der Entscheidungssituationen auswählen, wobei alle Entscheidungssituationen die gleiche Wahrscheinlichkeit besitzen ausgewählt zu werden. Der Computer bestimmt Ihre Auszahlung gegeben der von Ihnen gewählten Option (LINKS oder RECHTS) und der entsprechenden Wahrscheinlichkeitsverteilung. Dieses Ergebnis ist Ihre tatsächliche Auszahlung für Teil 1 und wird vom Computer zwischengespeichert. Falls Teil 1 am Ende des Experiments zufällig als auszahlungsrelevant bestimmt wird, erhalten Sie dieses Ergebnis ausgezahlt.

Bitte heben Sie die Hand, falls Sie noch weitere Fragen an die Studienleitung haben. Sollten Sie keine Fragen haben, dürfen Sie Teil 1 am Computer starten.

Anleitung Teil 2

Teil 2 des Experiments besteht aus 6 Entscheidungssituationen bezüglich unsicherer Lotterien. Jede Entscheidungssituation besteht aus einer Testphase von 20 Runden (nicht auszahlungsrelevant), gefolgt von einer Entscheidungsrunde (potentiell auszahlungsrelevant). Diese Aspekte werden nun im Detail erklärt.

Lotterien: In jeder Entscheidungssituation treffen Sie eine Entscheidung zwischen den Lotterien LINKS und RECHTS, ohne jedoch zu Beginn Informationen über die jeweiligen Wahrscheinlichkeitsverteilungen und die möglichen Auszahlungen zu haben. Diese Informationen können Sie in der Testphase sammeln.

Testphase: Die Testphase besteht aus 20 Runden in denen sie die Lotterien LINKS und RECHTS ausprobieren können, um so eine Stichprobe aus der zugrundeliegenden (unbekannten) Wahrscheinlichkeitsverteilung zu ziehen. Sie sehen sofort im Anschluss an jede Proberunden das Ergebnis Ihrer gewählten Lotterie. Die Stichproben sind unabhängig voneinander und werden mit Zurücklegen gezogen (d.h. immer aus der gleichen zugrundeliegenden Verteilung). Die Ergebnisse der Testphase sind nicht auszahlungsrelevant. Nach 20 Runden erhalten sie eine Mitteilung, dass die Testphase zu Ende ist.

Entscheidungsrunde: Die Entscheidungsrunde jeder Entscheidungssituation folgt direkt auf die jeweilige Testphase. In der Entscheidungsrunde treffen Sie einmalig auf Basis der in der Testphase gesammelten Informationen eine potentiell auszahlungsrelevante Entscheidung zwischen den Lotterien LINKS und RECHTS, aus denen Sie in der Testphase Stichproben gezogen haben.

Wie zuvor erwähnt kann eine der 6 Entscheidungssituationen aus Teil 2 für Ihre Auszahlung relevant sein. Nachdem Sie alle 6 Entscheidungen getroffen haben wird der Computer zufällig eine der Entscheidungssituationen auswählen, wobei alle Entscheidungssituationen die gleiche Wahrscheinlichkeit besitzen ausgewählt zu werden. Der Computer bestimmt Ihre Auszahlung gegeben der von Ihnen gewählten Option (LINKS oder RECHTS) und der entsprechenden Wahrscheinlichkeitsverteilung. Dieses Ergebnis ist Ihre tatsächliche Auszahlung für Teil 2 und wird vom Computer zwischengespeichert. Falls Teil 2 am Ende des Experiments zufällig als auszahlungsrelevant bestimmt wird, erhalten Sie dieses Ergebnis ausgezahlt.

Bitte heben Sie die Hand, falls Sie noch weitere Fragen an die Studienleitung haben. Sollten Sie keine Fragen haben, dürfen Sie Teil 2 am Computer starten.

1.G Experimental instructions - English translation

General Instructions

Dear participants,

Welcome and thank you for participating in our todays' experiment. Please, make sure that during the experiment, you:

- Do not talk with other participants.
- Switch off your mobile phone.
- Do not make any notes.

Should you have any questions during the experiment, please raise your hand. One of the experimenters will then come to your seat, to answer your question.

This experiment consists of two parts. In each part, you will make a decision in 6 different situations, henceforth called decision situations. At the end of the experiment, one of the two parts will be randomly selected. From this part, one decision situation will be randomly selected in turn and the outcome of your decision in this situation will be paid out to you. In addition, you will receive a safe payment of €3. Your total payoff at the end of the experiment thus amounts to:

Total payoff = Payoff of the randomly selected task of the randomly selected part + €3.

Before we start with the experiment, please examine the information sheet on uncertain lotteries and answer the comprehension question. Once you have done this, please raise your hand and one of the experimenters will subsequently check your answer. Afterwards, you will receive the instructions for Part 1 and can proceed with the experiment. Once you finished Part 1, please raise your hand again, in order to receive the instructions for Part 2. You will proceed through this experiment at your own speed, independent of the other participants.

The overall procedure of the experiment is depicted briefly on the back of these instructions.

Procedure

In each round, you will have to decide between LEFT and RIGHT. One randomly chosen decision situation from the two parts will determine your payoff at the end of the experiment.

General instructions + comprehension question

Raise your hand, to get the comprehension question checked and to get the instructions for Part 1.

Part 1

Raise your hand, to receive the instructions for Part 2.

Part 2

Random draws to determine the payoff relevant part and task.

Results screen

Fill in the receipt

Questionnaire

If you have questions, please raise your hand.

Instructions Part 1

Part 1 of the experiment consist of 6 decision situations, in which you have to decide between the lotteries LEFT and RIGHT. Both lotteries will result in a payoff that respectively depends on a probability distribution. You can see an example of this below. You will not encounter this example as a decision situation in the experiment; it serves for illustrative purposes only.

Example decision situation

LEFT		RIGHT	
Probability	Payoff	Probability	Payoff
50%	**€10.00**	**25%**	**€12.50**
25%	**€7,50**	**25%**	**€7.50**
25%	**€2.50**	**50%**	**€5.00**

Please, choose between LEFT and RIGHT

Would you choose LEFT in this example, you would receive €10 with a probability of 50%, €7.50 with a probability of 25%, and €2.50 with a probability of 25%. Would you choose RIGHT, you would receive €12.5 with a probability of 25%, €7.50 with a probability of 25% and €5 with a probability of 50%.

One of these 6 decision situations can be relevant for your payoff, as mentioned before. After you made all 6 decisions, the computer will randomly draw one of the decision situations. All decision situations have the same probability to be picked. The computer will then determine your payoff given your chosen option (LEFT or RIGHT) and the corresponding probability distribution. This result is your actual payoff for Part 1 and will be saved by the computer. If Part 1 is randomly selected at the end of the experiment to be payoff relevant, this result will be paid out to you.

Please raise your hand, if you have further questions for the experimenters. If you have no questions, you may begin with Part 1 at the computer.

Instructions Part 2

Part 2 of the experiment consists of 6 decision situations concerning uncertain lotteries. Each decision situation consists of testing stage of 20 rounds (not payoff relevant) and a subsequent decision round (potentially payoff relevant). These aspects will now be described in further detail.

Lotteries: In each decision situation, you will make a decision between the lotteries LEFT and RIGHT, without having initial information concerning the respective probability distribution and the possible outcomes. You can collect this information in the testing stage.

Testing stage: The testing stage consists of 20 rounds in which you can test the lotteries LEFT and RIGHT to draw a sample from the underlying (unknown) probability distribution. Immediately after each testing round, you will see the outcome of your chosen lottery. The samples are independent from each other and will be drawn with replacement (i.e. always from the same underlying probability distribution). The outcomes of the testing stage are not payoff relevant. After 20 rounds, you will receive an announcement that the testing stage is over.

Decision round: The decision round in each decision situation follows directly the respective testing stage. In the decision round you will make a singular, potentially payoff relevant decision between the lotteries LEFT and RIGHT based on the information collected from both lotteries in the testing stage.

One of the 6 decision situations of Part 2 can be relevant for your payoff, as mentioned before. After you made all 6 decisions, the computer will randomly draw one of the decision situations. All decision situations have the same probability to be picked. The computer will then determine your payoff given your chosen option (LEFT or RIGHT) and the corresponding probability distribution. This result is your actual payoff for Part 2 and will be saved by the computer. If Part 2 is randomly selected at the end of the experiment to be payoff relevant, this result will be paid out to you.

Please raise your hand, if you have further questions for the experimenters. If you have no questions, you may begin with Part 2 at the computer.

Chapter 2

Reverse Bayesianism: Revising Beliefs in Light of Unforeseen Events

Abstract[‡]

Bayesian updating is the dominant theory of learning. However, the theory is silent about how individuals react to events that were previously unforeseeable or unforeseen. We test if subjects update their beliefs according to "*reverse Bayesianism*", under which the relative likelihoods of prior beliefs remain unchanged after an unforeseen event materializes. Across two experiments we find that participants do not systematically deviate from reverse Bayesianism. However, we do find well-known violations of Bayesian updating. Furthermore, decision makers vary in their ex-ante unawareness depending on the context.

[‡]Joint work with Tigran Melkonyan, Eugenio Proto, Stefan T. Trautmann and Andis Sofianos.

The authors thank several colleagues for valuable input. In particular, Ala Avoyan, Andrew Caplin, Stefano Caria, David Cesarini, Cary Deck, Jürgen Eichberger, Guillaume Frechette, Peter Hammond, Andrea Isoni, Edi Karni, Christian König-Kersting, Fabian Paetzel, Joerg Oechssler, Daniela Puzzello, Andy Schotter, Daniel Sgroi, and Stefan Traub provided numerous insights. We also thank seminar and conference participants at the Indiana University, New York University, University of Sussex, the ESA 2020 Global, CESifo Area Conference on Behavioral Economics 2020, Annual Canadian Economics Association Meeting 2021, RUD 2021, and Experimental Finance 2021 for helpful comments. We are also very grateful to J. Philipp Reiss for the use of the Karlsruhe Decision & Design Lab. Both experiments were pre-registered in the AEA Registry. The University of Glasgow, University of Heidelberg, and University of Warwick provided funding for this research.

2.1 Introduction

We live in a world where scientific progress, human activities, and events outside of our control constantly lead to discoveries and observations of *unforeseen*[5] and *unforeseeable*[6] phenomena that fundamentally change our worldviews and behavior. Situations with unforeseen or unforeseeable events are abundant.[7] Even when we can imagine rough outlines of a phenomenon or even when we have a rather precise understanding of its characteristics, we often overlook it in the construction of our universe or include it in the description of the universe but render it as impossible. Examples of such phenomena include global pandemics, political and economic crises, as well as scientific groundbreaking discoveries.

In some cases, the distinction between unforeseeable and unforeseen is blurred and individual-specific. As a result of differences in knowledge and cognitive capacity, what is foreseeable for some people might be unforeseeable to others. A very notable example of an event that was foreseeable but unforeseen by many is COVID-19. Numerous scientists and observers, including Bill Gates, have repeatedly warned us about the possibility of a disastrous pandemic. However, these warnings have been largely ignored by many policy makers, public health officials, and economic decision-makers. The financial crisis of 2007 had a similar nature. For example, Lehman Brothers assumed in 2005 that the worst-case scenario in the housing market was a temporary price depreciation of 5% over the next three years, followed by a rebound and price increase of 5% thereafter; a scenario with a substantial drop in prices over an extended period was not even considered (Gennaioli & Shleifer, 2018, p. 52).

Given the empirical relevance of unawareness and neglected events, the current paper aims to provide insight into how people update their beliefs when unforeseen events materialize. In such cases, Bayes rule is silent about how individuals update their beliefs and is not useful in formulating subsequent reactions. A number of different approaches have been advanced to examine behavior under such circumstances. The epistemic and choice-theoretic approaches are the two main strands in the literature. The goal of the epistemic approach (e.g. Dekel et al., 1998; Modica & Rustichini, 1999; Heifetz et al., 2006; Halpern & Rêgo, 2008; Grant & Quiggin, 2013, among others) is to develop logical approaches and definitions of awareness, unawareness, and partial awareness in non-strategic and strategic settings. By its very own nature, this approach is concerned with laying the epistemic foundations of unawarneness rather than producing readily testable hypotheses how decision-makers perceive and react to unforeseen events. The choice-theoretic approach (e.g. Kochov, 2010; Ortoleva, 2012; Karni & Vierø, 2013; Schipper, 2013; Grant & Quiggin, 2015; Grant et al., 2017; Chambers & Hayashi, 2018; Dietrich, 2018; Dominiak & Tserenjigmid, 2021; Schipper, 2022, among others) develops repre-

[5] We call an event unforeseen if either (i) a decision-maker is aware that the event may occur but assigns zero probability to it or (ii) she is not aware that the event may occur, but in principle could be (based on available information).

[6] We call an event unforeseeable when the information available to a decision maker is objectively insufficient to allow her to contemplate the existence of the event.

[7] Related concepts are Knightian uncertainty and the “unknown unknowns”, a term famously coined by the late former US Secretary of Defense Donald Rumsfeld.

sentations of preferences in the presence of unawareness from behavioral axioms on individual preferences. A central property in the literature on decision making under growing awareness is *reverse Bayesianism* (Karni & Vierø, 2013, 2015, 2017; Karni et al., 2020), according to which decision-makers react to prior null events by proportionately shifting probability mass to these events from the prior non-null events. That is, a reverse Bayesian's construction of a new universe maintains consistency with the old structure. The centrality of reverse Bayesianism stems from a number of factors. First, it is normatively appealing in terms of how information is used to form beliefs. Second, many important models of exchangeable random partitions in statistics and combinatorial decision-theory (e.g. Schipper, 2022) as well as behavioral models of unawareness (e.g. Dominiak & Tserenjigmid, 2021; Piermont, 2021) are either consistent with reverse Bayesianism, or have a non-trivial overlap with it. Furthermore, it is often used as a yardstick against which models of updating beliefs under unawareness are compared. Reverse Bayesianism is also intuitively simple and amenable to testing using behavioral data. In light of all these considerations, we focus our exploration of behavior under unawareness on testing reverse Bayesianism.

Reverse Bayesianism imposes a rationality constraint on the process of updating beliefs following null events. Suppose, for example, a decision-maker bets repeatedly on the color of a randomly drawn marble from an urn which she beliefs to contain 50 black marbles and 50 white marbles. At one point the decision-maker witnesses some number (known or unknown) of red marbles being unexpectedly added to the urn. Under this design, the contents of the original two-color urn are part of the updated three-color urn. Put differently, the "old world" remains a part of the "new world." As the information about the old world did not change, a rational updating rule requires that the decision-maker's posterior beliefs put equal probability weight on white and black marbles, irrespective of the number of red marbles added to the urn. This is exactly the updating process reverse Bayesianism predicts.

This updating is, however, often less trivial than the above example might suggest. With skewed distributions of beliefs or multiple initial outcomes, keeping likelihood estimates proportional to each other becomes challenging. Complying with the demands of reverse Bayesianism might hence be cognitively demanding in many circumstances. Bayesian updating is commonly violated by decision makers in many situations for similar reasons (Tversky & Koehler, 1994; Sonnemann et al., 2013; Benjamin, 2019). In the context of unforeseen events, descriptive validity may be affected by the asymmetric impact of the new information on the evaluation of existing events. We will discuss different mechanisms for such an asymmetric impact in the next section, including asymmetric salience, $1/N$-bias, and hindsight bias. It follows from these considerations that whether, and if so, how well decision makers adhere to reverse Bayesianism, is not immediately clear.

The present paper develops two experiments to study the formation of beliefs under growing awareness, and more specifically, to determine whether individuals adhere to reverse Bayesianism. We design our experiments so that the new and unexpected environment retains some parts of the old world. In addition, we test how belief formation and updating are moderated by the environment of the decision situation. According to our knowledge, the present paper is the first to experimentally examine belief formation and reactions to unforeseen events. Furthermore, it is the first experimental study of how expectations of the unknown evolve as the universe expands. We find that behavior in both experiments is consistent with reverse Bayesianism, despite the fact that the participants exhibit some commonly observed judgment biases. Based on our findings, reverse Bayesianism seems to be a natural updating rule for decision-makers, being compelling both from a normative and descriptive perspective.

A controlled laboratory experiment is perhaps the only environment where it is possible to perform our empirical exercise. Unforeseeable events are rare, and by definition it is impossible to predict them and set the stage for observing beliefs in a sufficiently accurate way. At the same time, in the controlled environment of an experiment, it is virtually impossible to generate objectively unforeseeable events. Our experimental designs involve events that vary by the degree of foreseeability. To distinguish them from objectively unforeseeable and objectively foreseeable events, we coin the events in our experiments as *reasonably unforeseeable* and *reasonably foreseeable*. Whether an event belongs to one of these two latter categories depends on the amount of information received by a participant in the experiment. Our empirical analysis reveals that participants generally do not expect the unknown when it is reasonably unforeseeable. In contrast, some expect an unknown event when it is reasonably foreseeable.

In the first experiment, we analyze behavior of participants who face either a reasonably unforeseeable or a reasonably foreseeable event. In the course of the experiment, we elicit beliefs about the content of an urn as well as willingness to sell a gamble that pays according to a prize randomly drawn from the urn. In both treatments of the experiment, the task entails the introduction of a new urn (with new prize(s)), which was hidden from the participants, and the subsequent addition of its content to the original urn. We find strong evidence for reverse Bayesianism in both treatments. Prior to encountering the surprise in the form of a new urn, participants on average estimate that the probability of a yet unobserved prize is zero. This probability estimate also remains zero after witnessing the surprise, except for some of the participants that were forewarned about the possibility of new prizes in their treatment. We further investigate how the nature of the surprise affects beliefs and the valuation of prospects. Two patterns emerge. First, in the treatment with forewarning about new prizes, higher valuations of prospects are always observed. That is, the possibility of the unknown seems to instill hope, rather than fear. Second, valuations increase (decrease) after a positive (negative) surprise, showing that decision makers do incorporate the new information.

In the second experiment, the new events are all reasonably foreseeable. Participants explore a digital urn by sequentially extracting marbles from it with replacement. They receive no information on which colors are possible, so that the reasonably foreseeable event is represented by so far unobserved colors. This setup allows us to study how their beliefs evolve over time. Participants in the second experiment are found to suffer from common Bayesian updating violations. Nevertheless, we again find that beliefs are updated in accordance with reverse Bayesianism. Additionally, we find that participants lower their perceived likelihood of further unknown events as they sample more or observe more unforeseen colors. Despite this, beliefs about potentially unforeseen events are very persistent, and about one third of the participants still expect a yet unobserved color even after 30 draws.

The rest of the paper is organized as follows. In Section 2.2 we provide a theoretical framework and derive the hypotheses that are tested in both experiments. In Section 2.3 we present the design and results of Experiment 1. Section 2.4 is dedicated to Experiment 2. Section 2.5 provides a general discussion of the results and concludes the paper. In the appendix we include the experimental instructions and some additional analysis for each experiment.

2.2 Theoretical Background and Hypotheses

Following Karni & Vierø (2017), let A denote a finite, non-empty set of actions and C_0 denote a finite, non-empty set of *feasible consequences*. To illustrate this framework, consider a pharma company appraising which of two research programs to invest in. Both programs are aimed at developing a drug to treat certain medical condition Y. The set A is given by the two research programs, a and b, the pharma company is considering for investment. The set C_0 represents the consequences of its choice in terms of either developing an effective drug to treat Y or being unsuccessful in that endeavor, denoted by U. Thus, we have $C_0 = \{Y, U\}$ in our example. Let also $x = \neg C_0$ denote an abstract residual consequence, which stands for "something other than what the decision-maker can describe" – for example, finding a treatment for some other medical condition that the pharma company could find a treatment for, but which it is not aware of.

The sets $\hat{C}_0 = C_0 \cup \{x\}$ and A together define the *augmented conceivable state space* via $\hat{C}_0^A := \{s : A \to \hat{C}_0\}$. That is, the *augmented conceivable state space* takes into account the possibility that an action may lead to the "everything else" consequence x. Moreover, the space of *fully describable conceivable states* is defined as $C_0^A := \{s : A \to C_0\}$, where the mappings' image is restricted only to describable consequences.

The augmented conceivable state space can be expanded by observing a new consequence $c' \notin C_0$. In our example, the pharma company may subsequently realize that, as a third consequence, either research program may produce a drug that treats some alternative medical condition Z instead of medical condition Y. The set of feasible consequences then expands to $C_1 = C_0 \cup \{c'\}$. In our example, $C_1 = \{Y, Z, U\}$. Furthermore, C_1^A and $\hat{C}_1^A$ can be defined analogously to C_0^A and $\hat{C}_0^A$, respectively.

Denote π_0 and π_1 as probability measures defined on C_0^A and C_1^A, respectively, and representing beliefs before and after a new consequence is observed. In addition to standard axioms guaranteeing an expected utility representation, Karni & Vierø (2017) impose an axiom of invariant risk preferences and two awareness consistency axioms. The latter three axioms ensure that preferences for different levels of awareness are consistent with each other. The resulting representation is characterized by expected utility preferences for different levels of awareness and reverse Bayesianism, with the latter requiring that for all $s, s' \in C_0^A$:

$$\frac{\pi_0(s)}{\pi_0(s')} = \frac{\pi_1(s)}{\pi_1(s')}.$$

That is, this model implies that the decision maker will hold the ratio of probability estimates for known outcomes constant after observing an unforeseen outcome. Under classical Bayesian updating, new information shrinks the state space by excluding some outcomes that had previously been assigned a positive prior probability. In contrast, the present model focuses on the reverse situation, where new information can expand the state space, while still making sure that beliefs are updated in accordance with Bayes rule. Hence, the name “reverse Bayesianism”.

Once a new consequence is discovered, a decision-maker will update her beliefs about further possible new outcomes, now captured by $x = \neg C_1$. Observing a new outcome can have either an increasing or decreasing effect (or none) on the decision maker’s awareness about unforeseen events. On the one hand, it is possible that the discovery of new consequences decreases the amount of remaining unforeseen consequences. On the other hand, discovering a new consequence may highlight that there are still unforeseen consequences to uncover. Accordingly, the model allows for both a decrease or increase in the probability assigned to the residual consequence.

Based on the above framework, we will test if the normative reverse Bayesian model matches the actual behavior of participants in incentivized decision-making experiments. In our study, decision makers state their beliefs about the likelihoods of different events (prizes in Experiment 1 and colors of marbles in Experiment 2) and express their willingness to accept (WTA) for the prospects in Experiment 1, using standard incentivizing procedures. The descriptive validity of reverse Bayesianism in this context is not trivial, given the large literature on violations of Bayesian updating. In the context of unforeseen events, descriptive validity may be affected by the asymmetric impact of the new information on the evaluation of existing events (high versus low prizes in Experiment 1; frequent versus less frequent colored marbles in Experiment 2). Different mechanisms for an asymmetric impact are conceivable. First, experience of unforeseen events may lead to asymetric salience of different, previously observed prizes or colors. Second, $1/N$ bias may asymmetrically affect events considered more or less likely before observing the new outcome (Sonnemann et al., 2013). Third, hindsight bias has been shown to lead to revisions of ex-ante beliefs: this may loosen the connection between beliefs before and after a new outcome is observed, regarding the previously observed outcomes (Hoffrage & Gigerenzer, 2000).

For Experiment 1, the set of fully describable conceivable states is given by different combinations of prizes in an urn. The probability of an unknown prize is elicited implicitly and a participant’s likelihoods of already observed prizes are not restricted to sum to 1. For Experiment 2, the set of fully describable conceivable states is given by different combinations of colored marbles in an urn. The probability of an unknown event is elicited explicitly by asking participants to state their belief about “any other possible color.”

We differentiate between the *original urn* (before a new outcome is observed) and the *updated urn* (after a new outcome is observed). In the following we denote the probability estimates of each participant for a given state i by $\hat{p}_i^o$ and $\hat{p}_i^u$, for the original and the updated urn, respectively. The residual estimate is denoted by $\hat{p}_x^o$ and $\hat{p}_x^u$. Our two experiments will test the following main hypothesis:

Hypothesis 1: *Participants update their beliefs according to reverse Bayesianism. That is, for any $\hat{p}_i^o, \hat{p}_i^u$ and any states $i, i' \in C_0^A$:*

$$\frac{\hat{p}_i^o}{\hat{p}_{i'}^o} = \frac{\hat{p}_i^u}{\hat{p}_{i'}^u}.$$

In the analysis of the two experiments, we will refer to the difference between the ratios before and after the update as:

$$\Delta R = \frac{\hat{p}_i^u}{\hat{p}_{i'}^u} - \frac{\hat{p}_i^o}{\hat{p}_{i'}^o}.$$

In some of our experimental treatments, we explicitly rule out the possibility of unforeseen events and inform the participants about this. If a participant trusts that information, then $\{x\}$ should be empty for her and, as a result, $\hat{C}_0 = C_0$. Specifically, we will test the following:

Hypothesis 2: *At point t of the elicitation, the residual estimate:*

(a) $\hat{p}_x^t = 0$ *in the treatments where unforeseen events are ruled out so that $\{x\}$ is empty and $\hat{C}_0 = C_0$.*

(b) $\hat{p}_x^t > 0$ *in the treatments where unforeseen events are not ruled out.*

A further novelty of our experiments is that we explicitly study how residual estimates change after the state space expands. As outlined above, the framework is flexible in regard to how a decision maker's awareness reacts to a new event. Specifically, it allows for decision makers to increase their residual in either direction after observing a new, unforeseen event. Accordingly, we test the following agnostic hypothesis:

Hypothesis 3: *Participants will not adjust their residual belief after encountering a new event:*

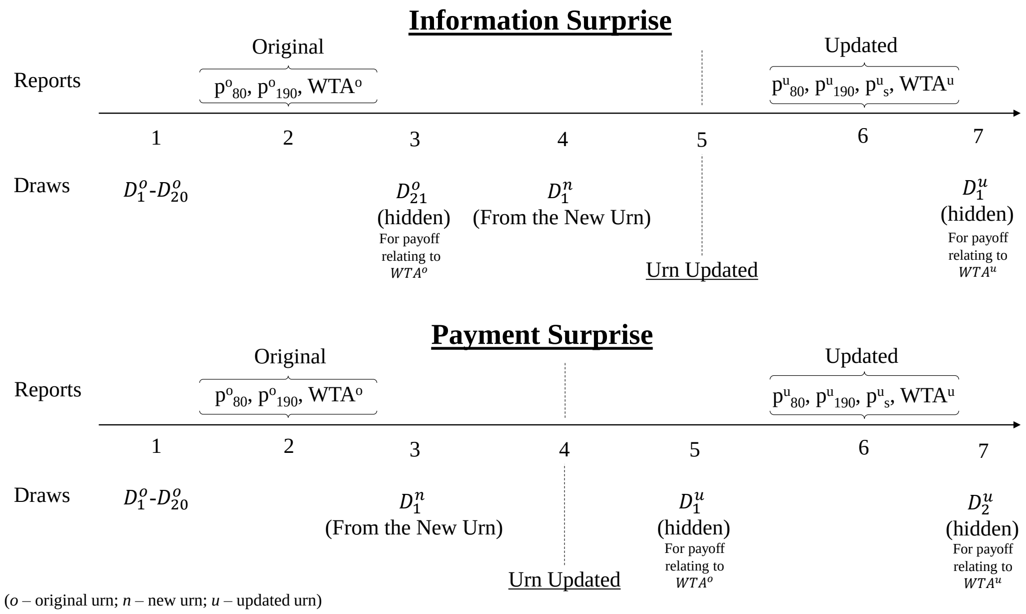
$$\hat{p}_x^u - \hat{p}_x^o = 0.$$

2.3 Experiment 1

2.3.1 Design

Experiment 1 elicits beliefs and valuations of prospects before and after encountering a new event. We test reverse Bayesianism using either a reasonably unforeseeable (*Information surprise/IS*) or reasonably foreseeable (*Payment surprise/PS*) event (we omit the word "reasonably" in this section). Each of these two conditions employs either a favorable or adverse new event (*high prize, low prize*), resulting in a 2 × 2 between-subjects design. Figure 2.1 provides an overview of the timing in the experiment. We will provide a rationale for our choice of the four conditions after we spell out their details. Our reasoning behind the labels *Information surprise* and *Payment surprise* will also become apparent.

Figure 2.1: The timing of the two surprise conditions.



Under the *IS* condition, the participants are presented with an urn, called the *original* urn, and are informed that the urn contains balls with labels representing prizes measured in tokens. Each earned token is exchanged for €0.05 at the end of the experiment. The participants are told that: "*the urn contains two and only two prizes*". However, they are not told what these two prizes or their relative proportions are. Furthermore, we do not alert the participants that the composition of the urn might change by adding or removing balls from the urn.

Following the description of the urn, the participants observe a sequence of 20 physical draws with replacement from the original urn ($D_1^o - D_{20}^o$ in Figure 2.1). The original urn contains 24 balls labeled '80' and 36 balls labeled '190'. No information regarding the specific composition of the urn is disclosed to the participants. All of the draws are made by a participant we refer to as the "experimental assistant," who is randomly selected for this task from the participants in a given session. The outcome of each draw is revealed to all participants by the experimental assistant.[8] Thus, everybody in a session observes the same sample. None of these 20 draws are payment-relevant. The only purpose of these draws is for the participants to gain information about the composition of the original urn.

After observing these 20 draws from the original urn, the participants are asked to provide estimates of the probabilities of the prizes they have been observing during these draws (subjective probabilities p_{80}^o and p_{190}^o) and to state their willingness-to-accept (minimum selling price) to sell the prospect of drawing a prize from this urn (WTA^o). We use superscript 'o' in our

[8]The experimental assistants do not complete any of the tasks that the other participants perform and receive a fixed payment of €14, which is close to the average earnings in the experiment.

notation to emphasize that these values are elicited before any changes to the *original* urn are made. The reported estimates of the two probabilities do not have to add up to 100%. The design does not force the sum of the estimates to be lower or greater than 100% either. We do not, however, explicitly ask for the respondents' estimates of observing a prize that they have not observed during the 20 draws. Thus, our design allows for an implicit estimation of a residual probability of outcomes that have not yet been observed. This contrasts with Experiment 2 where we explicitly ask for that probability.

Following the 20 draws and the reports of the two probabilities and WTA, a draw from the original urn is made by the experimental assistant (D^o_{21} in Figure 2.1). The outcome of this draw determines the potential payments for the reports of WTA^o. However, the draw is concealed from the participants when it is made. It is revealed only at the very end of the experiment when the final payment to the participants is displayed, provided that this decision is selected for payment.

Subsequently, we bring out a new urn to the front of the experimental lab. One ball is then drawn from the new urn (D^n_1 in Figure 2.1), revealing a new prize s to all participants. At this point in the experiment, the participants are informed that: "*This urn contains only the prize you are (about to be) shown.*" The value s of the new prize varies with the prize condition. In the *low* prize condition, the new urn contains 15 balls labeled '15', while in the *high* prize condition, it contains 15 balls labeled '375'. Although the participants know the value of all prizes in the new urn, they are not informed about how many balls are contained in the new urn. After revealing the value of the prizes in the new urn, we empty its contents into the original urn.[9] We call this combined urn the *updated* urn. The participants are then asked to estimate the probabilities of each of the three prizes ($p^u_{80}, p^u_{190}, p^u_s$) and to state their WTA for the prospect to draw a prize from the updated urn (WTA^u). We use superscript 'u' in our notation to emphasize that these values are elicited after the urn is updated. Following these reports, the experimental assistant draws a ball from the updated urn (D^u_1 in Figure 2.1). This draw is concealed and is only revealed at the very end of the experiment, provided this decision is selected for payment.

Consider now the *PS* condition. Similarly to the *IS* condition, the participants are presented with the *original* urn, and are informed that the urn contains balls with labels that represent prizes. Again, each earned token is exchanged for €0.05 at the end of the experiment. In contrast to *IS*, the participants are not told about the number of different prizes in the *original* urn. Similarly to *IS*, they are not provided with any information about the proportions of balls with any specific prize. In the *PS* condition the respondents are informed that: "*at any point in the study new balls representing different tokens to what you have been observing so far may be added to this urn*". Thus, one might expect that some participants may incorporate this piece of information, which is not provided under *IS*, into their process of arriving at and reporting

[9] To prevent the respondents from inferring the number of balls in either urn from the first 20 draws and emptying of the new urn into the original one, all of the balls were made of styrofoam while the boxes were made of opaque plastic material.

probabilities and WTAs. Similarly to *IS*, the participants subsequently observe 20 physical draws from the original urn ($D_1^o - D_{20}^o$ in Figure 2.1).

After observing the 20 draws from the *original* urn, the participants are asked to report their estimates of the probabilities of the prizes that they have been observing (subjective probabilities p_{80}^o and p_{190}^o) and to state their willingness-to-accept to sell the prospect of drawing a prize from the urn (WTA^o). As for *IS*, the reported estimates of the probabilities are not restricted to add up to 100% or to be smaller or larger than 100%, allowing for calculation of an implicit residual probability.

Following the elicitation of these probabilities and WTA^o, we bring out a new urn to the front of the experimental lab. The participants are informed that: "*This urn contains new prizes. One such prize is the one you see. The urn contains no prizes similar to what you have been observing as a result of random draws from the other urn.*" The experimental assistant subsequently brings in the new urn and draws one ball from it (D_1^n in Figure 2.1), revealing one new prize s to the participants. We do not reveal any other information about the contents of the new urn. As before, the value of the new prize s varies with the prize condition of the session. In the *low* prize condition, the new urn contains 15 balls labeled '15', while in the *high* prize condition it contains 15 balls labeled '375'. We then proceed to empty the contents of the new urn into the original urn, leading to the *updated* urn. The experimental assistant subsequently makes a draw from the updated urn (D_1^u in Figure 2.1). The outcome of the draw from the updated urn is used to determine the potential payment for the report of WTA^o. However, the draw is concealed from the participants immediately after it is made. As for *IS*, it is revealed only at the very end of the experiment when the final payment to the participants is displayed, provided that the WTA^o report is selected for payment. Thus, in contrast to the *IS* condition where the draw determining payment for the report of WTA^o is made before the urn is updated, the draw in the *PS* condition is made from the updated urn: participants were forewarned that this may happen.

The participants are then again asked to estimate the probability of each prize ($p_{80}^u, p_{190}^u, p_s^u$) and to state their willingness-to-accept for the prospect to draw a prize from the urn (WTA^u). Following these reports, the experimental assistant draws a ball from the updated urn (D_2^u in Figure 2.1). This draw is used to determine the payment to the respondents, provided the WTA^u report is selected for payment at the very end of the experiment. Similarly to *IS*, this draw is concealed from the respondents and only revealed at the very end of the experiment if this decision is selected for payment. Thus, the last two stages of the *PS* condition coincide with the last two stages of the *IS* condition (stages 6 and 7 in Figure 2.1).

The payment for the urn tasks is determined as follows. One item is randomly selected from the set of all reported probability estimates and the two WTAs. This item is played out and the resultant payoff is added to a participant's payment. We incentivize the reported probability estimates according to the Karni (2009) method.[10] The reports of WTA are incentivized using the BDM procedure (Becker et al., 1964). Both mechanisms induce truth telling and are robust to varying risk attitudes.

The BDM procedure works as follows: If one of the WTAs is selected to be payment-relevant, the computer draws a random price for the prospect between the smallest and largest prizes in the urn.[11] If the realization of the random price is such that the reported WTA exceeds the randomly generated price, the participant keeps the prospect and the payoff is determined by the hidden draw made during the experiment from the original urn in *IS* condition for WTA^o (see D^o_{21} in Figure 2.1); from the updated urn for WTA^u (see D^u_1 in Figure 2.1); and from the updated urn for both WTA^o and WTA^u in *PS* condition (see D^u_1 and D^u_2 in Figure 2.1). If WTA is smaller than the random price, the participant sells the prospect, and receives the random price. The Karni mechanism works similarly for probabilities. A random probability (between 0 and 1) is drawn. If the reported probability estimate exceeds this randomly drawn probability, the participant is rewarded with the relevant prize according to the actual probability of that prize in the urn and gets nothing with the complementary probability. Alternatively, the participant is rewarded with the relevant prize according to the randomly drawn probability and gets nothing with the complementary probability.

The objective of our study is to examine whether subjects expect the unexpected and to investigate how their beliefs change after encountering a new event. For a new event to be unexpected, it should be unannounced and/or ruled out. To elicit updated beliefs within our design in an incentive compatible fashion, the new event must have direct and immediate payment consequences. Ideally, our treatment would have both of these characteristics. However, if our individual conditions had both of these characteristics, we could be accused of deception for explicitly or implicitly signaling that there would be no new event but then implementing the latter and making it payment-relevant. In light of this constraint, we designed our experiment to have two conditions, each of which has one and only one of these characteristics. Notice that the aim of the two conditions, *IS* and *PS*, is not to contrast belief updating directly between the two, but rather, to study belief updating in two closely related situations. In *IS*, the new event is unannounced while in *PS* the new event is payment-relevant because the payment is determined by the updated urn. Specifically, in contrast to *IS*, the respondents in *PS* are forewarned that new prizes may be added to the urn (thus, making the new event potentially foreseeable). Moreover, while in *IS* the first (potentially) payment-relevant draw (D^o_{21} in Figure 2.1) is made from the original urn, the corresponding draw in *PS* is made from the updated urn (D^u_1 in Figure 2.1). Our approach also allows us to test the robustness of results regarding reverse Bayesianism across different settings.

In addition to these two differences, the conditions *IS* and *PS* differ along two other dimensions. These two differences were implemented to make the new event in *IS* as unexpected as

[10]See the first page of the experimental instructions in Appendix 2.A following the heading "Likelihoods of events – Reporting and Earnings" for more details on how this was explained to the participants as well as for further details on the method itself.

[11]Similarly to Isoni et al. (2011), these bounds are not communicated to the participants.

possible and to avoid misleading the respondents in *PS* as much as possible. These differences pertain to the information about the compositions of the original and new urns, respectively, under the two conditions. Unlike in the *IS* condition, we did not tell the respondents in the *PS* condition that the original urn contains two and only two prizes. Otherwise, one could argue that we are sending a message that contradicts the possibility that the content of the urn may be changed or that we are trying to mislead the respondents. Even after making 20 draws from the original urn, the respondents in *PS* may expect to encounter a prize value that they have not yet observed. Thus, the possibility of drawing a new prize is conceivable in *PS*. Finally, the fourth difference is consonant with the third. The respondents in *IS* are told that the new urn contains only balls with the newly revealed prize (D_1^n in Figure 2.1). In contrast, under *PS* the possibility that the new urn may contain prizes other than the newly revealed prize is not ruled out.

Once the urn task is completed, we elicit risk preferences using an incentivized Eckel & Grossman (2008) task. In this task, individuals pick one lottery from a set of binary lotteries. The lotteries in the choice set vary in terms of their expected values and variances with the chosen lottery revealing a participant's risk attitude. The lotteries chosen by the participants are "played out" at the end of the experiment and the earnings for these choices are added to the rest of the earnings of each participant.

Following the elicitation of risk preferences, the participants complete a short Raven Advanced Progressive Matrices (APM) test (Raven et al., 1998b,a). Raven's Progressive Matrices provide an effective non-verbal avenue to measure reasoning and general cognitive ability. In order to shorten the duration of this test, we follow Bors & Stokes (1998) in using 12 from the total of 36 matrices from Set II of the APM. Matrices from Set II of the APM are appropriate for adults and adolescents of higher than average intelligence. Participants are allowed a maximum of 10 minutes. The participants are informed that two of these 12 matrices are selected at random for payment and that they will receive €1 for each correct choice. The sessions are concluded with some general demographic questions and a final screen informing the participants about their total earnings.

We include the experimental instructions in Appendix 2.A. Within this appendix in table 2.5, we also include a table that lists the draw realizations across all sessions and the average WTA reported by session. The design was pre-registered at the AEA RCT Registry `https://www.socialscienceregistry.org/trials/3815`.

Implementation

Experiment 1 was conducted at the Alfred-Weber-Institute Experimental Lab at the University of Heidelberg and the Karlsruhe Decision & Design Lab (KD^2Lab) at the Karlsruhe Institute of Technology. Conducting the experiment in two separate locations was done to reduce the possibility that former participants communicate details about the experiment to later participants, and given the nature of this experiment it would have been particularly problematic. The recruitment of participants took place via SONA systems for Heidelberg and ORSEE (Greiner, 2015) for Karlsruhe. A total of 344 participants participated in the experimental sessions.[12] The participants earned an average of €18.4, including a show-up fee of €4. The software used for the entire experiment was z-Tree (Fischbacher, 2007). The ethical approval for this design was granted by the Humanities and Social Sciences Research Ethics Sub-Co at the University of Warwick under DRAW Umbrella Approval (Ref: HSS 49/18-19, DR@W submission ID: 485613261).

2.3.2 Results

Reverse Bayesianism

We start by testing whether belief updating is consistent with reverse Bayesianism. In our framework, reverse Bayesianism requires that the elicited probability ratios of the prizes in the original urn, namely the prizes of 80 and 190 tokens, remain unchanged after the original urn is updated. Formally, Hypothesis 1 for this experiment implies:

$$\Delta R = \frac{\hat{p}^u_{80}}{\hat{p}^u_{190}} - \frac{\hat{p}^o_{80}}{\hat{p}^o_{190}} = 0. \tag{2.1}$$

The information provided to the participants in all four treatments unambiguously reveals that the number of balls worth 80 tokens and the number of balls worth 190 tokens remain unchanged after the urn is updated. For (*IS, low prize*) and (*IS, high prize*) treatments, we informed the participants that the new urn contains only the newly revealed prize. For (*PS, low prize*) and (*PS, high prize*) treatments, we told them that the new urn contains no prizes similar to what they have been observing as a result of random draws from the original urn. Thus, if a participant's beliefs are given by a singleton probability distribution, i.e., she is probabilistically sophisticated (Machina and Schmeidler, 1992), throughout the experiment, then $\Delta R = 0$ as long as that participant forms her beliefs on all of the information provided in the experimental instructions.

Table 2.1 contains the results of Wilcoxon signed-rank tests for all four treatments. We fail to reject the hypothesis in three out of the four treatments, both before and after correcting for multiple testing. For these three treatments, we indeed find a precisely estimated null effect.[13] Only in one treatment, (*PS, high prize*), we reject the null hypothesis of reverse Bayesianism. Looking at the confidence intervals derived from a t-test, we note that 95% of participants deviate very little from 0, even in the treatment where we reject the null hypothesis. Subjects seem to change the probability ratio in the updated urn by decreasing $\hat{p}_{190}$ slightly more than $\hat{p}_{80}$. Since

[12] We had 234 participants in the sessions at Heidelberg and 110 participants in the sessions at Karlsruhe. In Heidelberg, 46 participants were in the (*IS, low prize*) treatment, 58 in (*IS, high prize*), 59 in (*PS, low prize*), and 71 in (*IS, high prize*). In Karlsruhe, the numbers were 30, 17, 34, and 29, respectively. Data is not qualitatively different across the two subject pools.

[13] Testing the ratios before and after the update also shows that participants do not simply provide equal estimates for both prizes, that is having ratios equal to 1. On average, ratios before and after the update are smaller than 1, with $p < 0.001$, Wilcoxon signed-rank test.

the original urn contains a larger number of balls worth 190 tokens than balls worth 80 tokens, this adjustment of reported probabilities might arise from subjects feeling more comfortable with decreasing a larger number. We report a similar behavioral pattern in experiment 2.

It is important to highlight that our data analysis so far fails to reject the null hypothesis not because of limited power, but because the null is supported by our data. To provide further evidence in support for the predicted null effect of reverse Bayesianism, we rely on Bayesian inference statistics.[14] Specifically, we implement the JZS test developed by Rouder et al. (2009), which is a Bayesian alternative to t-tests. Like other Bayesian methods it offers a researcher the possibility to state whether the data contains evidence in support of the null hypothesis. Crucially, the JZS test makes assumptions about the prior distributions of the effect size and variance, thus circumventing the problem of Bayes factors favoring the null hypothesis when a non-informative prior is used for the alternative.[15] The usual rule-of-thumb for interpreting Bayes factors applied to the JZS test is that a factor of 0 provides strong evidence for the alternative, 1 is inconclusive (the predictions of the null and the alternative cannot be disentangled) and values above 3 provide strong evidence for the null hypothesis (here reverse Bayesianism). The last column in Table 2.1 reports the Bayes factors. These values are consistently above 3 in all four treatments, thus offering strong evidence in support of reverse Bayesianism.[16]

Table 2.1: Average ratio changes before and after the urn is updated.

		Obs	Avg ratio change	p-value	p-value (corr)	95% CI	Bayes factor
IS	low prize	75	0.007	0.375	1.000	$[-0.06, 0.05]$	14.76
	high prize	75	-0.039	0.981	1.000	$[-0.06, 0.14]$	6.57
PS	low prize	93	0.016	0.918	1.000	$[-0.06, 0.03]$	9.72
	high prize	100	-0.007	0.011	0.043	$[-0.04, 0.05]$	16.35

Notes: Wilcoxon signed-rank test, p-values corrected by Bonferroni-Holm procedure, confidence interval from one sample t-test, Bayes factor from JZS test.

The ratio of the probabilities of the prizes observed by the participants may remain unchanged under two scenarios. First, participants may simply not change their estimate for any of the previously observed prizes. Table 2.2, column 7, shows that this is rarely the case in any of the treatments. Second, participants may change their estimates in a way that keeps the ratios constant. Notice that this is far from trivial for many constellations of estimates. Although we observe such instances (column 4), overall this is considerably less prevalent than the instances

[14] For discussions and examples of the use of Bayesian inference statistics in the social sciences, see, e.g., Bayarri et al. (2016); Dienes (2011).

[15] Specifically, the JZS test assumes that the null hypothesis is a point with $H_0 : \delta = 0$, while the alternative effect size $\delta = \frac{\mu}{\sigma}$ follows a Cauchy distribution (which assumes that more extreme values are more unlikely) with $H_1 : \delta \sim Cauchy(r) = 1$ (where r is a scaling parameter). The prior for the variance is given by $p(\sigma^2) \propto \frac{1}{\sigma^2}$. This prior is deliberately non-informative, as the variance is relevant for both hypotheses.

[16] Juxtaposing the estimated confidence interval and the Bayes factors for the (*PS, high prize*) treatment suggests that the rejection of the null according to the Wilcoxon signed-rank test is possibly driven by a relatively large proportion of positive changes that are nevertheless of very small magnitude. This is further supported by the bottom right panel of Figure A2.3 in Appendix 2.C.

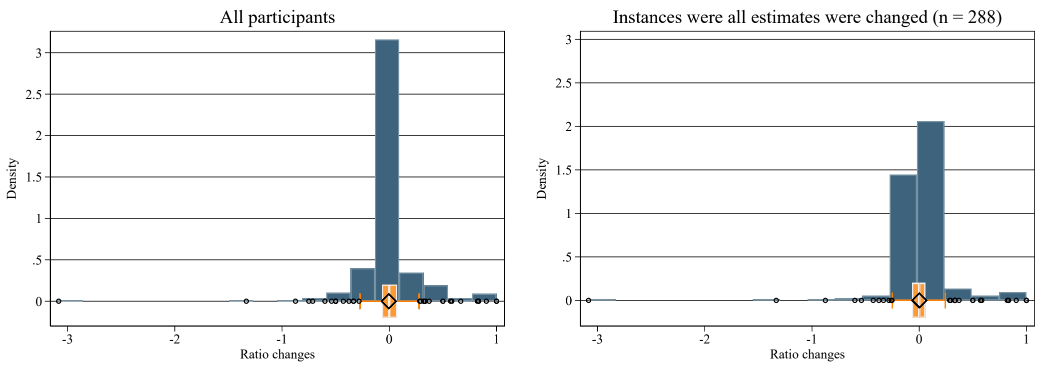
when participants either increased or decreased the ratio. Figure 2.2 shows the distribution of ratio changes both for all participants and for all instances where the participants adjusted all their estimates after the update. The data is pooled over all four treatments.[17] Our analysis reveals that the absolute differences of the ratios tend to be concentrated around zero. That is, the change in the ratio is relatively small even for those participants who updated their beliefs to a different ratio. However even when the ratio changes, as we already noted from Table 2.2, most participants who change their ratios do so only slightly, and in no systematic direction. Thus, the overall null effect in support of reverse Bayesianism materializes because the changes in the ratio of the probabilities tend to be very small.

Table 2.2: Changes of the probability ratios following the update of the urn.

		Increased	Decreased	Const ratio	p-value	p-value (corr)	Unchanged Est
IS	low prize	29	23	23	0.488	1.000	1
	high prize	31	32	12	1.000	1.000	1
PS	low prize	33	37	23	0.720	1.000	0
	high prize	29	61	10	0.001	0.004	4

Notes: Matched pairs sign test, p-values corrected by Bonferroni-Holm procedure. 'Unchanged Est.' denotes the subset of those holding their ratios constant while not changing any of their estimates.

Figure 2.2: Histograms of the changes in the ratios following the update of the urn.



Notes: Histogram in blue, box plot in orange, outliers (circles) and mean (diamond) in black.

The evidence we have provided so far indicates that on average we observe no changes to the ratios. It is important to emphasize that this does not mean that the participants are passive and do not update their estimates after the original urn is updated. In Table 2.3, we test whether the estimates for known outcomes are significantly different between the original and updated urns. Specifically, we compare individual estimates of the prizes before and after the update. We consistently find that the estimates of the known outcomes are significantly higher in the original urn as compared to the updated urn. This indicates that the participants do react to

[17]See Figure A2.3 in Appendix 2.C for the equivalent distributions for each treatment separately.

the updating of the urn when reporting their proportion estimates. Summarizing:

Result 1: *The participants hold their ratios approximately constant after encountering an unexpected event. Thus, they update their beliefs according to reverse Bayesianism (Hypothesis 1).*

Moreover, we do not find a statistically significant relation between cognitive ability and ratio differences (see Table A2.2 in Appendix 2.C). That is, behavior is very similar for high and low cognitive ability participants, thus, unlikely to be due to errors caused by lack of understanding.

Result 2: *Cognitive ability has no mediating effect on the deviations from reverse Bayesianism.*

Table 2.3: Changes of known outcome estimates after observing the update of the urn.

		Obs	Diff	p-value	p-value (corr)
IS, low prize	$\hat{p}_{80}^u - \hat{p}_{80}^o$	76	-0.101	0.000	0.000
	$\hat{p}_1^u 90 - \hat{p}_1^o 90$	76	-0.130	0.000	0.000
IS, high prize	$\hat{p}_{80}^u - \hat{p}_{80}^o$	75	-0.102	0.000	0.000
	$\hat{p}_1^u 90 - \hat{p}_1^o 90$	75	-0.125	0.000	0.000
PS, low prize	$\hat{p}_{80}^u - \hat{p}_{80}^o$	93	-0.100	0.000	0.000
	$\hat{p}_1^u 90 - \hat{p}_1^o 90$	93	-0.136	0.000	0.000
PS, high prize	$\hat{p}_{80}^u - \hat{p}_{80}^o$	100	-0.075	0.000	0.000
	$\hat{p}_1^u 90 - \hat{p}_1^o 90$	100	-0.108	0.000	0.000

Notes: Wilcoxon signed-rank test, p-values corrected by Bonferroni-Holm procedure.

Residuals and Valuations

The experiment yields a number of additional interesting findings. First, we focus on the estimates of residual probabilities provided by the participants and test Hypotheses 2 and 3. Turning to Hypothesis 2, the participants in the *IS* condition were informed that the original urn contains only two possible prizes, and that in the updated urn only one extra prize was added. If the respondents took this information into account, the set of residual consequences x, as defined in Section 2.2, would be empty, thus $\hat{p}_x^o \neq 0$ and $\hat{p}_x^u \neq 0$ would likely only materialize due to individual idiosyncratic errors. In contrast, since in the *PS* condition we informed the participants about the possibility of adding new prizes and we did not state that the updated urn contains only one new prize, $\hat{p}_x^o > 0$ and $\hat{p}_x^u > 0$ could occur as a result of an expectation of unknown events. Thus, one could expect that the respondents in the *PS* condition are more likely to assign a strictly positive probability to encountering a prize that they have not seen before.

Table 2.4 shows that the hypothesis $\hat{p}_x^t = 0$ (t is either o for original or u for updated) cannot be rejected for the *IS* condition, both for the original and for the updated urn, thus giving support to Hypothesis 2 (see also Figures A2.1 and A2.2 in Appendix 2.C for a graphical representation of the residual estimates). Even in the *PS* condition, the hypothesis $\hat{p}_x^t = 0$

cannot be rejected for the original urn. Moreover, the hypothesis cannot be rejected in the *PS* condition for the updated urn in the case of (a perhaps less salient) high prize surprise. The hypothesis that $\hat{p}_x^t = 0$ is rejected in the updated urn in the (*PS, low prize*) treatment. Moreover, column 2 of Table 2.4 reveals that in the *PS* condition, $\hat{p}_x^u > 0$ for a larger number of participants compared to other conditions. It seems plausible that at the point in the experiment where the participants have only seen the original urn, the event that the urn may be updated is unforeseeable to many of them. In contrast, the latter event is unlikely to be unforeseeable after the participants have witnessed the update of the original urn.

Result 3: *Overall, the participants do not expect unforeseen events, thus, lending support to the first part of Hypothesis 2. Evidence on the second part of Hypothesis 2 is mixed: after encountering adverse new events, participants anticipate unforeseen events, thus, supporting the second part. However, after favorable unforeseen events, they do not, thus rejecting the second part.*

Table 2.4: Residuals different from 0.

		$\hat{p}_x^t = 0$	$\hat{p}_x^t > 0$	$\hat{p}_x^t < 0$	p-value	p-value (corr)
IS, original	low prize	74	1	1	0.993	1.000
	high prize	71	3	1	0.314	1.000
PS, original	low prize	92	0	1	0.317	1.000
	high prize	90	6	4	0.549	1.000
IS, updated	low prize	61	10	5	0.251	1.000
	high prize	65	7	3	0.228	1.000
PS, updated	low prize	74	16	3	0.004	0.028
	high prize	84	11	5	0.146	1.000

Notes: Wilcoxon signed-rank test, p-values corrected by Bonferroni-Holm procedure.

We now turn to Hypothesis 3. Following the same reasoning as before, we expect some awareness for *PS* but not for *IS*, and test for changes in the residuals after receiving new information. Table 2.5 shows that the hypothesis that $\hat{p}_x^o = \hat{p}_x^u$ cannot be rejected in the *IS* treatments. However, there is again some evidence for $\hat{p}_x^o(PS) \neq \hat{p}_x^u(PS)$, when the surprising event entails a low prize.

Result 4: *With the exception of adverse new event for* PS *condition, the participants do not adjust their residual beliefs following an unforeseen event, thus, not rejecting Hypothesis 3.*

Given our design, awareness of encountering an unforeseen event and the way an unforeseen event is experienced will affect the stated $WTAs$. It follows from Karni & Vierø (2013, 2017) that two key factors are at play in the participants' evaluation of uncertain prospects. The first characteristic pertains to the participants' updated beliefs; how much of the probability weight is shifted from the known prizes to the newly observed and yet unobserved prizes. The second concerns the participants' attitude towards the unknown; whether and how much they like or dislike the unknown. To determine the relative importance of these channels, we compare the

Table 2.5: Differences between the residual before and after the surprise: $\Delta\hat{p}_x = \hat{p}_x^u - \hat{p}_x^o$.

		$\Delta\hat{p}_x = 0$	$\Delta\hat{p}_x > 0$	$\Delta\hat{p}_x < 0$	p-value	p-value (corr)
IS	low prize	60	11	5	0.173	0.692
	high prize	63	6	6	0.937	1.000
PS	low prize	73	17	3	0.002	0.009
	high prize	82	11	7	0.345	1.000

Notes: Wilcoxon signed-rank test, p-values corrected by Bonferroni-Holm procedure.

elicited willingness to accept measures before and after the urns are updated, for both the *IS* and *PS* conditions and both levels of the new prize. Table 2.6 reveals that $WTA^o(PS) > WTA^o(IS)$ in both high and low prize conditions. As one could expect, the $WTAs$ for the updated urn are lower for the low prize than for the high prize, and again the *PS* condition elicits higher valuations; the latter effect is, however, not significant.

The regression analysis with controls for gender, cognitive ability, degree of risk aversion and the observed sample, confirms the effect of the *PS* treatment and of the high prize in the updated urn *WTA* (see Tables A2.3 and A2.4 in Appendix 2.C). Overall, it seems that the more uncertain situation in condition *PS* elicits higher valuations. That is, in the context of unforeseen events, *hope* seems to dominate *fear* (Viscusi & Chesson, 1999). Additionally, it seems that the belief about the number of 190 prizes in the urn is an important driver of the WTA, not the actually observed number. As a caveat, we note that WTA measurement in the context of uncertainty and ambiguity has been found to elicit relatively higher valuations for more uncertain prospects (Trautmann et al., 2011; Trautmann & Schmidt, 2012). The selling-price context seems to induce decision makers to focus on the potentially forgone benefits from selling a highly uncertain prospect. This effect re-emerges here.

Result 5: *The increased uncertainty in condition PS results in higher valuations of the urn: the participants appear to view the unknown with hope rather than with fear.*

2.4 Experiment 2

2.4.1 Design

Experiment 2 tests how individuals perceive uncertainty and update their beliefs in light of new events which are foreseeable, but potentially unforeseen. Each participant in Experiment 2 individually draws a sample of 30 colored marbles with replacement from a virtual urn in each of four different tasks. The urns in each of the four tasks contain a total of 100 virtual marbles, which is known to the participants. No other information about the composition of the urns is revealed to the participants prior to them making the draws from the urns. There are two types of potentially unforeseen events in this design. The first type entails encountering a new color, while the second pertains to observing some *specific* new color. The way the task is set up, we

Table 2.6: *WTA* for a draw from the urn by treatment.

Original urn: WTA^o

	IS	PS	Diff	p-value	p-value (corr)
Low prize	110.39	138.47	-28.08	0.008	0.031
High prize	110.48	134.81	-24.33	0.002	0.006

Notes: Wilcoxon signed-rank test, p-values corrected by Bonferroni-Holm procedure.

Updated urn: WTA^u

	IS	PS	Diff	p-value	p-value (corr)
Low prize	86.45	96.70	-10.25	0.074	0.295
High prize	153.53	178.25	-24.72	0.160	0.639

Notes: Wilcoxon signed-rank test, p-values corrected by Bonferroni-Holm procedure.

expect participants to assign a non-zero belief to the first type of event. This, however, does not mean that participants cannot be surprised in this experiment. The belief participants assign to specific events of the second type could vary considerably between participants, depending on their imagination (Shackle, 1949). Hence, while we expect participants to be aware about the existence of new colors in general, they should be surprised by *specific* colors unforeseen by them. In our investigation, we focus on eliciting beliefs about the first type of events, which is also the union of all possible events of the second type.[18]

Each participant is randomly allocated to one of two treatments. In the *two colors* treatment, the urn in the first task contains only two colors. In the *four colors* treatment, the urn in the first task contains four colors in total. The purpose of this design is to test if encountering a larger number of different colors in the first task increases their awareness that further surprises, in the form of new colors, might be possible in subsequent tasks. The compositions of the urns in the second, third and fourth tasks are the same across the two treatments. The urn in the second task contains three colors, the urn in the third task contains two colors and the urn in the fourth task contains four colors. Table 2.7 provides information on the exact compositions of the urns in the four tasks.

The urns with three and four colors contained a comparatively small number of some of the colors. This ensured that the likelihood of encountering a new and surprising outcome even after sampling several times was relatively large for these urns.

After each sample draw, the drawn marble is presented on a participant's screen, both with an image of a marble of that specific color and the name of the color. In addition, the outcomes of all previous draws are depicted in a small caption at the bottom of the participants' screens. This provides the participants with a full overview of the past draws and mitigates the effects

[18]Exact colors for each task are randomized at the participant level. As a result, each participant observed a different sequence of colored marbles for each task.

Table 2.7: Numbers of different colors in the tasks.

	Task 1 Two colors	Task 1 Four colors	Task 2	Task 3	Task 4
Color 1	55	40	53	75	48
Color 2	45	28	35	25	28
Color 3		20	12		12
Color 4		12			12

of memory limitations.

After each draw, the participants are asked to report their estimates of the contents of the urn. Specifically, they are asked to state their estimate of (i) the number of marbles of the color they *just drew*; (ii) the number of marbles of *each other* color they have previously drawn in this task; and (iii) the number of marbles of *yet unobserved* colors.

The third item provides us with a residual probability assigned by the respondents to any conceivable color not yet observed during the draws. Thus, in contrast to Experiment 1, the residual probability is elicited explicitly in Experiment 2. The participants submit their estimates by entering integer numbers between 0 and 100 into form fields on their screens. We provide them with one individual form field for each color drawn up to that point, as well as with one form field for their estimate of the number of marbles of yet unobserved colors. This design allows us to trace how the respondents' estimates and ratios of these estimates are adjusted once unforeseen information becomes available. In addition, we can track how the estimates of likelihoods of yet unobserved outcomes evolve over the sampling process. To make the submission of estimates easier, we additionally provide respondents with buttons that allow them to fill in their last estimates for a color and "plus" and "minus" buttons to increase/decrease these estimates by one per click.[19]

We again use the Karni (2009) method to ensure that the participants are incentivized to submit their estimates truthfully. At the end of the experiment, for each of the four tasks, one of the 30 sample draws is randomly selected and one item from the set of reported estimates for that draw is also randomly chosen by the computer (this could involve an estimate of yet unobserved colors). The payment mechanism is implemented for that reported estimate. Participants can earn £6 or nothing from each task depending on the outcome according to the incentivization method.[20]

Following the four sample tasks, the participants complete the same incentivized APM task described used in Experiment 1 (see Section 2.3.1). After the APM, participants are informed in detail about their total earnings. The sessions are concluded with a questionnaire containing demographic questions and a question eliciting risk attitudes.

[19] An example screenshot of the sampling screen is included in Appendix 2.E.

[20] See the third page of the instructions for this experiment in Appendix 2.D under the heading "Getting paid for good predictions" for information on how this was explained to the participants as well as for further details on the method itself.

In addition to the hypotheses presented in Section 2.2, we additionally test the following:

Hypothesis 4:

(a) The probability $\hat{p}_x^t$ assigned each round to yet unobserved outcomes decreases with the draws (t) made from an urn:

$$t \to 30 \Rightarrow \hat{p}_x^t \downarrow$$

(b) The probability assigned to yet unobserved outcomes in tasks 2, 3 and 4 decreases faster for the participants in the two colors treatment than for the participants in the four colors treatment.

Hypothesis 4 tests if the participants learn to expect fewer surprises towards the end of the sampling process, as well as whether this process might vary in terms of speed across treatment conditions. Specifically, participants in the *four colors* treatment might be more aware that more colors are possible.

Following some recent insights on the link between cognitive ability and rational behavior (e.g. Alaoui & Penta, 2016; Gill & Prowse, 2016) we also test two hypotheses related to cognitive ability.

Hypothesis 5: *The participants with higher cognitive ability exhibit fewer deviations from reverse-Bayesianism.*

Hypothesis 6: *The participants with higher cognitive ability expect yet unobserved outcomes up to a later point in the sampling process than the participants with lower cognitive ability.*

We include the experimental instructions in Appendix 2.D. The design was pre-registered at the AEA RCT Registry `https://www.socialscienceregistry.org/trials/5499`.

Implementation

Experiment 2 took place at the Behavioral Science Lab at the University of Warwick. The recruitment was conducted with the DRAW (Decision Research at Warwick) system, based on the SONA systems. A total of 174 individuals participated in the experiment, 89 in the *two colors* treatment and 85 in the *four colors* treatment. Note that we originally intended to have 150 participants in each treatment (as specified in our pre-registration). However, due to the unforeseen onset of the COVID-19 pandemic we were not able to gather additional data. The average payment was £16.87, including a show-up fee of £3. The software for the experiment was programmed in otree (Chen et al., 2016). Ethical Approval for this design was granted by the Humanities and Social Sciences Research Ethics Sub-Co at the University of Warwick under DRAW Umbrella Approval (Ref: HSSREC 104/19-20, DR@W submission ID: 514470520).

2.4.2 Results

Reverse Bayesianism

We first test whether behavior is consistent with reverse Bayesianism, which corresponds to Hypothesis 1. We examine the difference in the ratios of previously observed colors directly before and directly after observing a new color. The ratios are defined for pairs of colors that have already been observed and on the basis of the relative magnitudes of the estimated likelihoods of these two colors immediately before a new color is observed. Specifically, we define $\hat{p}^o_H$ as the estimate of the likelihood for the color that is considered by a participant to be more likely and $\hat{p}^o_L$ as the estimate for the color that is considered to be less likely. Both of these estimates are for the sample draw right before the third color is observed for the first time. We also define $\hat{p}^u_H$ and $\hat{p}^u_L$ as the estimates of the likelihoods of these two colors immediately after the third color is first observed. Specifically, we test:

$$\Delta R^3 = \frac{\hat{p}^u_H}{\hat{p}^u_L} - \frac{\hat{p}^o_H}{\hat{p}^o_L} = 0.$$

For the belief update following the observation of a fourth color (having already seen three colors), we have three ratios to consider. Define $\hat{p}^o_H$, $\hat{p}^o_M$ and $\hat{p}^o_L$ as the estimates for the color considered most likely, second most likely, and least likely, of the three colors that have already been observed, in the sample draw right before the fourth color is observed for the first time. We also let $\hat{p}^u_H$, $\hat{p}^u_M$ and $\hat{p}^u_L$ denote the respective estimates for these three colors in the sample draw immediately after the fourth color is first observed. We test three relationships:

$$\Delta R^4_1 = \frac{\hat{p}^u_H}{\hat{p}^u_M} - \frac{\hat{p}^o_H}{\hat{p}^o_M} = 0, \quad \Delta R^4_2 = \frac{\hat{p}^u_M}{\hat{p}^u_L} - \frac{\hat{p}^o_M}{\hat{p}^o_L} = 0, \quad \Delta R^4_3 = \frac{\hat{p}^u_H}{\hat{p}^u_L} - \frac{\hat{p}^o_H}{\hat{p}^o_L} = 0.$$

Table 2.8 contains the results of Wilcoxon signed-rank tests for all ratio changes (as described just above) including the data from both treatments.[21] For Tasks 1 and 4, we also pool the three ratio changes after observing the fourth outcome (indicated in Table 2.8 as ΔR^4_P). The results indicate that for all eleven tests that we conduct, there is no significant change in the ratios after controlling for multiple tests.[22] Even when not controlling for multiple testing, there is no statistically significant change in eight of the eleven tested ratios.[23] Finally, from the 95% confidence intervals (obtained from t-tests) in column 5, we notice that the ratios do not change substantially. This is especially manifest when the urn contains three colors so that the participants have to keep only one ratio constant, denoted as ΔR^3 in the table. Similarly to the analysis for Experiment 1, we also include the Bayes factors in the last column of Table 2.8. Except for three cases, the Bayes factors are above 3. In two of the three cases, however, the confidence intervals still include 0. Table 2.8 thus provides strong evidence in support of reverse Bayesianism.

[21] Our results did not differ across treatments, thus our tables pool treatments from here on.

[22] The urn in task 3 contains only two colors. Hence, the third outcome surprise is not possible and, consequently, we do not analyze it in this case.

[23] As in Experiment 1, we also test whether participants before and after the update simply provide equal estimates for both prizes, that is having ratios equal to 1. On average, ratios before and after the update are larger than 1, with $p < 0.001$, Wilcoxon signed-rank test.

Table 2.8: Average ratio changes before vs. after observing a new color.

		Obs	Avg ratio change	p-value	p-value (corr)	$95\%CI$	Bayes factor
Task 1	ΔR^3	85	-1.365	0.172	1.000	$[-0.10, 0.29]$	5.32
	ΔR_1^4	84	-0.548	0.584	1.000	$[-0.79, 0.22]$	4.44
	ΔR_2^4	84	-2.134	0.033	0.362	$[-1.27, 0.04]$	1.58
	ΔR_3^4	84	-1.005	0.315	1.000	$[-0.52, 0.33]$	7.52
Pooled	ΔR_P^4	252	-2.229	0.026	0.284	$[-0.64, -0.03]$	1.49
Task 2	ΔR^3	169	-2.632	0.008	0.093	$[-0.31, 0.01]$	2.26
Task 4	ΔR^3	173	-0.648	0.517	1.000	$[-0.26, 0.06]$	5.71
	ΔR_1^4	164	-0.048	0.962	1.000	$[-0.07, 0.25]$	6.05
	ΔR_2^4	163	-0.067	0.946	1.000	$[-0.19, 0.69]$	6.09
	ΔR_3^4	163	-0.148	0.883	1.000	$[-0.14, 0.27]$	9.46
Pooled	ΔR_P^4	490	-0.203	0.839	1.000	$[-0.03, 0.30]$	5.64

Notes: Wilcoxon signed-rank test, p-values corrected by Bonferroni-Holm procedure, confidence interval from one sample t-test, Bayes factor from JZS test.

It is worth noticing that all average ratio changes, albeit statistically insignificant, are negative. This suggests the possibility of a minor pattern that is not captured because of the variability in the data. As in experiment 1, subjects might be decreasing the likelihoods of events to which they originally assigned relatively high probabilities by relatively large amounts (in this case, $\hat{p}_H$ decreases slightly more than $\hat{p}_M$ that, in turn, decreases slightly more than $\hat{p}_L$). This effect, if it exists, is small in magnitude and it remains insignificant after correcting for multiple testing and even when we pool together all of the changes, including the changes following the fourth event (i.e. ΔR_P^4). Furthermore, the fact that the absolute value ΔR_P^4 in task 1 is larger than in task 4 hints at a potential learning effect, where the participants' behavior aligns more with reverse Bayesianism in later tasks.

As noted in the analysis for Experiment 1, the finding that the ratios do not significantly change in aggregate may derive from (i) participants holding their estimates unchanged, (ii) participants holding the ratios constant, or (iii) some participants increasing while some decreasing their ratios, with the effects canceling each other out on average. Table 2.9 indicates that, in contrast to Experiment 1, a large number of participants hold their estimates unchanged. A possible explanation for the participants in Experiment 2 not changing their estimates is the provision of a button to fill-in their previous estimate to simplify the dynamic task for the participants. However, there is also a substantial number of participants who hold their ratio constant, while adjusting the separate probability estimates. Again, holding the ratio constant while adjusting the separate estimates, especially after the fourth color is observed, is far from trivial. It indicates that participants actively aim to keep their ratios constant even when changing their separate estimates. At the same time, a substantial share of the participants change their ratios, but there is no systematic effect in the way they do it: after correcting for multiple testing, there are no significant differences in the deviations from constant ratios in the increasing or decreasing directions; seven of the nine uncorrected tests support constant average ratios.[24]

Table 2.9: Changes of the ratios before vs. after observing a new outcome.

		Increased	Decreased	Const ratio	p-value	p-value (corr)	Unchanged Est
Task 1	ΔR^3	16	29	40	0.072	0.797	26
	ΔR_1^4	19	21	44	0.875	1.000	31
	ΔR_2^4	16	31	37	0.040	0.440	32
	ΔR_3^4	16	23	45	0.337	1.000	35
Pooled	ΔR_P^4	51	75	126	0.040	0.440	93
Task 2	ΔR^3	35	59	75	0.017	0.189	46
Task 4	ΔR^3	45	50	78	0.682	1.000	44
	ΔR_1^4	50	57	57	0.562	1.000	36
	ΔR_2^4	54	60	49	0.640	1.000	33
	ΔR_3^4	43	47	73	0.752	1.000	37
Pooled	ΔR_P^4	147	164	179	0.364	1.000	108

Notes: Matched pairs sign test, p-values corrected by Bonferroni-Holm procedure.'Unchanged Est' denotes the subset of those holding their ratios constant while not changing any of their estimates.

This is further corroborated by Table 2.10, where we test whether the participants update their estimates of known outcomes after observing a new color. We again compare individual estimates of known outcomes before and after observing a new color. Similarly to Experiment 1, even though the ratios are on average constant, the individual estimates of known outcomes are updated downwards as new colors are observed.

Figure 2.3 illustrates that, as in Experiment 1, the ratio changes were closely concentrated around zero. This again holds both when we pool all participants in the experiment and when we pool all instances where all estimates are changed after observing a new color.

Result 6: *The participants update their beliefs according to reverse Bayesianism, thus, providing support for Hypothesis 1.*

Once more, cognitive ability does not have a significant effect on the ratio deviations (Table A2.5 in Appendix 2.F). Thus, there is no empirical evidence supporting Hypothesis 5. Importantly, the absolute changes in the ratios are also unaffected by the participants' expectations of further surprises, that is, whether they hold non-zero residual beliefs or not (Table A2.6 in Appendix 2.F).

Result 7: *Cognitive ability has no mediating effect on deviations from reverse Bayesianism, thus, we reject Hypothesis 5.*

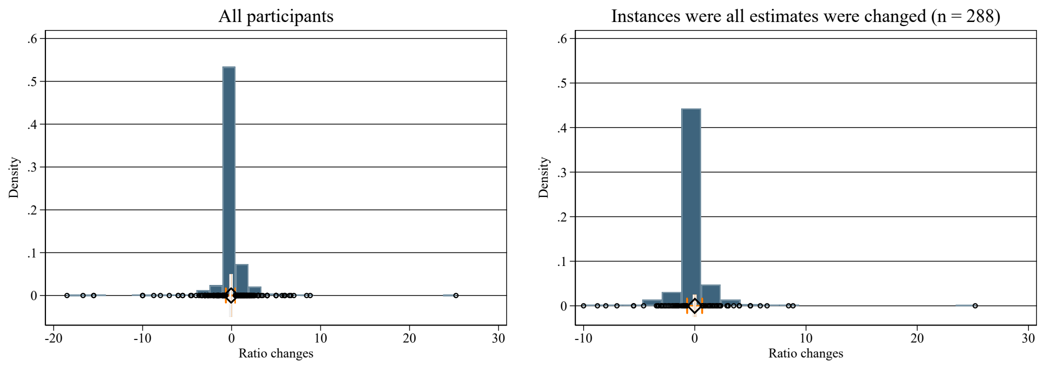
[24] Again, we also find the above described (albeit still insignificant) pattern for the pooled ratios ΔR_P^4 in Tasks 1 and 4.

Table 2.10: Changes of known outcome estimates after observing a new color.

		Obs	Diff	p-value	p-value (corr)
Task 1, after third color	$\hat{p}^u_H - \hat{p}^o_H$	85	-0.06	0.000	0.000
	$\hat{p}^u_L - \hat{p}^o_L$	85	-0.04	0.000	0.000
Task 1, after fourth color	$\hat{p}^u_H - \hat{p}^o_H$	84	-0.04	0.000	0.000
	$\hat{p}^u_M - \hat{p}^o_M$	84	-0.02	0.000	0.005
	$\hat{p}^u_L - \hat{p}^o_L$	84	-0.02	0.000	0.000
Task 2, after third color	$\hat{p}^u_H - \hat{p}^o_H$	169	-0.07	0.000	0.000
	$\hat{p}^u_L - \hat{p}^o_L$	169	-0.05	0.000	0.000
Task 4, after third color	$\hat{p}^u_H - \hat{p}^o_H$	174	-0.07	0.000	0.000
	$\hat{p}^u_L - \hat{p}^o_L$	174	-0.07	0.000	0.000
Task 4, after fourth color	$\hat{p}^u_H - \hat{p}^o_H$	164	-0.05	0.000	0.000
	$\hat{p}^u_M - \hat{p}^o_M$	164	-0.03	0.000	0.000
	$\hat{p}^u_L - \hat{p}^o_L$	164	-0.03	0.000	0.000

Notes: Wilcoxon signed-rank test, p-values corrected by Bonferroni-Holm procedure.

Figure 2.3: Histograms of the changes in the ratios following the update of the urn.



Notes: Histogram in blue, box plot in orange, outliers (circles) and mean (diamond) in black.

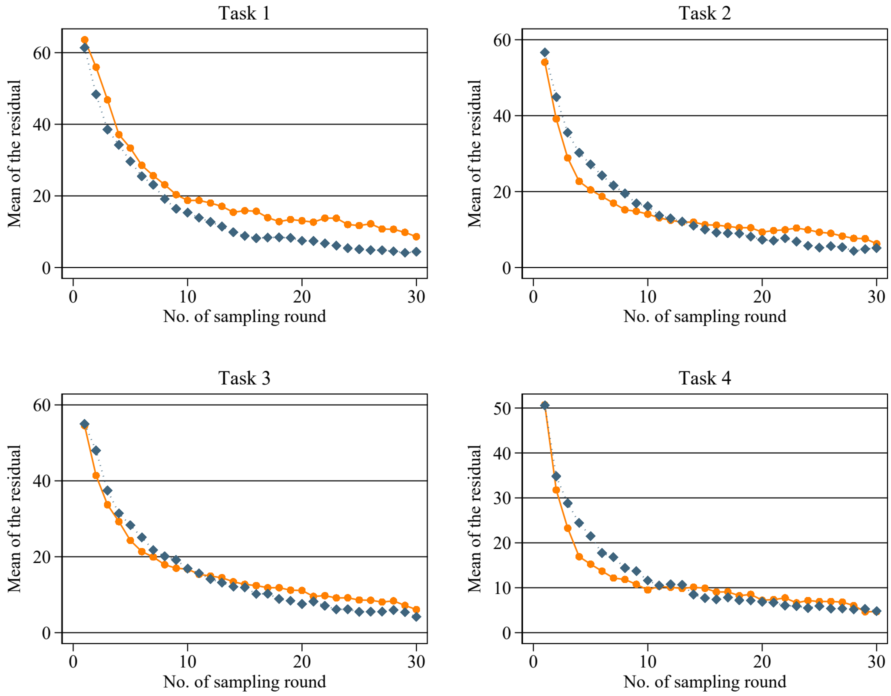
Analysis of Residuals and Belief Dynamics

We now turn to the estimates of the residual probabilities. Figure 2.4 depicts the evolution of the average residual $\hat{p}^t_x$ over the 30 sample draws for each task and treatment. We find that for both treatments the average residual starts at a relatively high level and decreases quickly as more draws are made. However, even after the 30^{th} sample draw, the average residual is well above 0. In Figure A2.5 in Appendix 2.F, we present the distribution of these residuals; about one third of the participants expect further colors even after the 30^{th} sample draw.

Result 8: *On average, the participants anticipate unforeseen events, $\hat{p}^t_x > 0$, thus, providing support for Hypothesis 2.*

For the sake of exposition we postpone the analysis of Hypothesis 3, and consider Hypothesis 4 first. We already noted that the residuals on average monotonically decrease with sample draws

Figure 2.4: Dynamics of the residuals over the sampling process.



Note: The orange line depicts the residuals of *two outcomes* treatment while the blue line depicts the residuals of *four outcomes* treatment.

(Figure 2.4). Simple correlations (across the whole set of participants) between the number of marbles already drawn from the virtual urn and the stated residuals are negative for all tasks and across both treatments ($\rho < -0.311$, Pearson correlation coefficient), and thus support the first part of Hypothesis 4. However, there are no significant differences between the two treatments in the overall shape of the curve (Kolmogorov-Smirnov test, all $p-values \geq 0.994$), thus rejecting the second part of Hypothesis 4 (treatment differences in awareness of encountering new colors). This indicates that participants do not adapt their estimation of residuals in later tasks in response to encountering more possible outcomes in the first task. This finding is intriguing. On the one hand, encountering more possible outcomes in the first task could raise the participants' awareness that the urns might contain more outcomes than initially expected, which was our prediction. On the other hand, as the participants have no information ex ante and as task 1 does not provide direct information on subsequent tasks, the null effect that we find might be perfectly rational.

Result 9: *The residual probabilities decrease with the draws made from the urn, thus, supporting Hypothesis 4a. There is no difference in this trend between the two and four colors treatments, thus, we reject Hypothesis 4b.*

Turning to Hypothesis 3 (that the residual probability changes after observing new colors), there is a negative correlation between the number of colors already observed by the participants and their residual probabilities ($r_s < -0.272$, Spearman correlation coefficient). Examining the changes in the residuals directly before and after a new color is observed reveals the same picture. Table 2.11 depicts the residuals after each update, with $\hat{p}_x^2$ denoting the residual after the second color is observed, $\hat{p}_x^3$ after the third, and $\hat{p}_x^4$ after the fourth. There tends to be a significant drop in the subjective residual probability in 8 out of 9 instances. Thus, the more colors a participant already encountered the smaller is her expectation of a new color.

Table 2.11: Changes of the residuals before vs. after observing a new color.

		Increased	Decreased	Constant	p-value	p-value (corr)
Task 1	$\hat{p}_x^2$	9	144	21	0.000	0.000
	$\hat{p}_x^3$	16	56	13	0.000	0.000
	$\hat{p}_x^4$	25	38	21	0.130	1.000
Task 2	$\hat{p}_x^2$	3	149	22	0.000	0.000
	$\hat{p}_x^3$	26	84	59	0.000	0.000
Task 3	$\hat{p}_x^2$	3	149	22	0.000	0.000
Task 4	$\hat{p}_x^2$	6	143	25	0.000	0.000
	$\hat{p}_x^3$	27	94	53	0.000	0.000
	$\hat{p}_x^4$	27	57	80	0.001	0.013

Notes: Matched pairs sign test, p-values corrected by Bonferroni-Holm procedure.

To summarize our findings pertaining to the residual probabilities reported up to this point: (1) drawing a larger sample decreases the residual as more precise information on already observed colors is available; (2) observing more colors, ceteris paribus, decreases the residual probability, perhaps because participants feel that the space of so far unobserved events shrinks with the number of observed colors.

In order to better understand to which degree these two factors impact the residual, we estimate a random effects model of the residual on the number of draws and the number of observed colors, controlling for different demographic factors and the task. Table 2.12 presents the results of this analysis. Both factors independently have a negative and significant impact on the size of the residual. The impact of the number of outcomes is roughly 11 times stronger than the effect of the sample draws. It is interesting to note that a higher cognitive ability leads to a smaller residual, in contradiction to Hypothesis 6. However, since it is not obvious what "optimal" residual an individual participant should have in this experiment, it is not possible to assess whether it is reasonable to observe this relationship. Finally, we run two robustness checks of the estimations presented in Table 2.12. First, the results pertaining to the coefficients and their significance remain robust when a fixed-effects instead of a random-effects model is used (Table A2.7 in Appendix 2.F). Second, conducting the random effects panel analysis solely upon the latter half of the sampling process (only for observations after sampling round 15), leads to a smaller yet still significant negative coefficient for the number of draws (Table A2.8 in Appendix 2.F). The coefficient for the number of colors observed becomes insignificant, possibly due to lower power with a relatively small number of new colors being observed in later rounds of the sampling process. Taken together, these results indicate that even in the later stages of the sampling process, participants on average reduce their residuals with every additional draw.

Result 10: *The participants consistently update their residual probabilities downwards, thus, rejecting Hypothesis 3.*

Result 11: *The participants of higher cognitive ability report smaller residual probabilities, thus, rejecting Hypothesis 6.*

Table 2.12: Random Effects Estimator: relation between the sample draws and residuals, panel GLS.

Size of the residuals	
Num. draws	-0.746**
	(0.048)
Num. colours observed	-8.402**
	(0.537)
Cognitive ability	-2.624**
	(0.547)
Four colours first	-0.914
	(2.509)
Constant	77.907**
	(11.212)
Observations	19,800
Subjects	165

Notes: * $p < 0.05$; ** $p < 0.01$
Standard errors in parentheses, standard errors are clustered at the individual level. The estimation additionally controls for age, gender, being an economics student and risk aversion but the coefficients are not reported.

Bayesian and Reverse Bayesian Rationality

As shown previously, the participants' behavior is consistent with reverse Bayesian reasoning. Does this mean that our participants are in general very rational Bayesian updaters? In order to test this, we study how observing a new color affects the sum of the new residual probability and the estimate of the new color. Technically, before actually observing a new color, its estimate should be included in the estimate of the event *any other color*. Observing a new color can be viewed as unpacking the estimate of the likelihood of yet unobserved colors into two new estimates, an estimate for the new color and another estimate for the yet unobserved colors. That is, in the absence of updating about the joint event, the sum of the estimate of the new color and the new residual should equal the previous residual. Tversky & Koehler (1994) and Sonnemann et al. (2013) find that the sum of such two unpacked estimates violates this principle in a context without learning; the sum of the unpacked estimates often exceeds the 'packed' estimate.

We study such unpacking effects in our data. We define $\hat{p}^u_{Sum}$ as the sum of the new residual and the estimate for the newly observed color C, i.e. $\hat{p}^u_{Sum} = \hat{p}^u_x + \hat{p}^u_C$. As discussed, $\hat{p}^u_{Sum}$ should equal the previous-round residual $\hat{p}^o_x$ if there is no learning. In general, early updates should imply larger learning effects. Thus, the learning component of $\hat{p}^u_{Sum}$ should decrease with the number of draws conducted. Table 2.13 tests if the ratios $\frac{\hat{p}^u_{Sum}}{\hat{p}^o_x}$ are different from 1. This is strongly the case for all instances of the updates. There is also a significant positive correlation between the number of colors observed and the unpacking ratio (Kendall's $\tau_A = 0.667, \tau_B = 0.785, p = 0.0095$). That is, the unpacking effects get more pronounced as more colors are observed, suggesting a substantial psychological unpacking effect in the spirit of Tversky & Koehler (1994), rather than a rational learning effect. Figure 2.5 illustrates the size of the effect for the color observed fourth. The estimate of the new color is virtually added to the previous estimate of the event *any other color*, keeping the latter estimate close to constant.[25] This suggests that our participants are indeed prone to violations of rational updating principles in the current context. As we have argued above, several factors may lead participants to violate the reverse Bayesian principles. This makes the strong evidence for reverse Bayesianism all the more remarkable.

Result 12: *The participants succumb to the 'unpacking' violation of rationality. Thus, behavior consistent with reverse Bayesianism is not a part of uniform adherence to the principles of Bayesian updating.*

Table 2.13: Unpacking the residual after observing a new color.

	Second colour	Third colour	Fourth colour
Task 1	1.23	1.46	2.32
Task 2	1.29	1.65	
Task 3	1.48		
Task 4	1.25	2.02	2.23

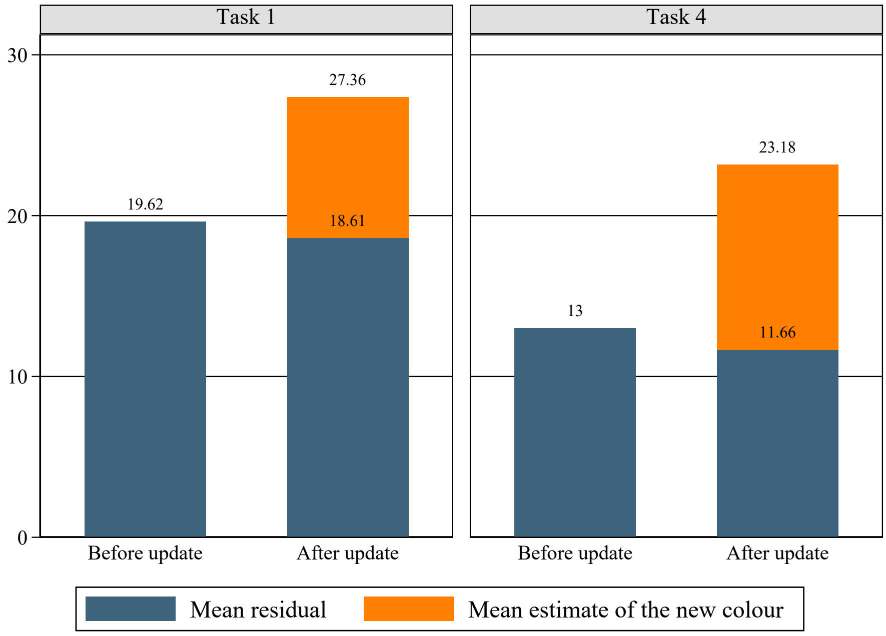
Notes: Wilcoxon signed-rank test, p-values corrected by Bonferroni-Holm procedure, all $p < 0.001$.

2.5 Concluding Remarks

There is a large literature assessing how decision makers update beliefs about known events in empirical and experimental decision situations (e.g., Charness & Levin, 2005; Charness et al., 2007; Grether, 1992; Holt, 2009). We focus on new and more or less unforeseeable events. Different strands of the theoretical literature offer varying prescriptions on how to integrate

[25] See Figures A2.6 and A2.7 in Appendix 2.F for an illustration of this when the second and third colors are observed.

Figure 2.5: Residuals and estimates of a new color before and after observing a new color, fourth color.



information about unforeseen events into beliefs. In their seminal work, Karni and Vierø (2013; 2017) axiomatize a preference functional of a decision maker who integrates unforeseen events into an updated probability distribution over an expanded state space so that the ratio of previously observed events stays unchanged. Our results provide evidence that reverse Bayesianism is also compelling from a descriptive perspective. This stands in sharp contrast to many studies on Bayesian updating, which often find behavior in the lab violating theoretical prescriptions (Charness & Levin, 2005; Charness et al., 2007; Holt, 2009). In other words, our results suggest that reverse Bayesianism is compelling, both, *normatively* and *descriptively*. This holds true both in situations involving reasonably unforeseeable events (Experiment 1) and in situations with unknown but foreseeable events (Experiment 2).

These implications are intriguing. The space of possible events is continuously expanding in many decision environments. Examples like the financial crisis of 2007 and the COVID-19 pandemic highlight the difficulty to react to novel events appropriately. It is thus important to pay special attention to the possibility of not knowing relevant events beforehand. While our experiments can not speak to the optimal reactions towards an objectively unforeseeable event, they show that decision makers have the capacity to reconcile new information optimally with their existing beliefs: encountering unforeseen events does not lead to a rearrangement of the "old" world of previously known events. Notably, this is irrespective of whether participants started with not explicitly expecting a surprise (Experiment 1) or if they were asked about such a belief and supplied a positive estimate (Experiment 2). Furthermore, it is noteworthy that the participants in Experiment 2 reduced their residual as more draws were made. This indicates an inclination towards becoming more complacent during the experiment, as participants lower their expectations of surprises as fewer new outcomes are discovered. Testing this tendency in more applied scenarios and studying how it interacts with precautionary measures could be an interesting extension to our findings.

Our findings for Experiment 1 also indicate that participants exhibit a higher WTA when they are aware of the possibility of further, unknown surprises. In the interpretation of Viscusi & Chesson (1999), the respondents in our study seem to be more hopeful rather than fearful towards unknown future events. This is also interesting in light of the ambiguity literature (Trautmann & van de Kuilen, 2015). One interpretation is that unforeseen events are assigned small probabilities (see also Experiment 2), and that the observed behavior is an embodiment of ambiguity seeking. Indeed, the Trautmann & van de Kuilen (2015) review reports predominantly ambiguity seeking behavior for low probability gain prospects like the ones used in Experiment 1.

In Experiment 2 (and implicitly in Experiment 1) we used the residual probability estimate as a catch-all way of encoding the participants' beliefs about all remaining possible events. This does not, however, give us a clear description of what exactly participants expect in the future, instead it is a "(...) Black Box, a residual of unknown content." (Shackle, 1992, p. 23). For example, a participant revealing a positive residual could be expecting a surprise in the future. Say she could be expecting the urn to also contain blue and red marbles. Observing a purple marble could present her with an unforeseen event. In our design, we do not elicit the precise nature of the unforeseen event - whether it corresponds to purple or some other specific color. Further studies could try to elicit an exhaustive list of expected events from the participants or even try to use a completely non-distributional approach to assess uncertainty (Shackle, 1992).

Furthermore, the surprises in our experiments might still be considered to be easily comprehensible. A new urn with new prizes and a new color in a box of colored marbles can be surprising and unexpected, but it is relatively simple to integrate their materialization into an existing belief structure after their first occurrence. A natural next step could be to study whether the principle of reverse Bayesianism extends to more complex events, where it is less clear which form a surprise might take and, importantatly, how to reconcile it with existing beliefs. This could help us to better understand how belief systems are affected by very rare and novel events.

Finally, our results cannot resolve the question whether participants act as-if they are reverse Bayesian or if their underlying thought process adheres to the prescriptions of reverse Bayesianism. On the one hand, some participants did not alter their estimates at all after observing a new outcome. This could suggest the as-if interpretation. On the other hand, a significant share of the participants did, and still provided belief ratios that were either constant or deviated little from the previous ratio. This would be more consistent with decision makers consciously following reverse Bayesianism. More research is needed to further disentangle these two possibilities.

Chapter 2 References

Alaoui, L., & Penta, A. (2016). Endogenous depth of reasoning. *The Review of Economic Studies*, *83*(4), 1297–1333.

Bayarri, M., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, *72*, 90–103.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, *9*(3), 226–232.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, *2*, 69–186.

Bors, D. A., & Stokes, T. L. (1998, jun). Raven's advanced progressive matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*, *58*(3), 382–398.

Chambers, C. P., & Hayashi, T. (2018). Reverse Bayesianism: A comment. *American Economic Journal: Microeconomics*, *10*(1), 315–24.

Charness, G., Karni, E., & Levin, D. (2007). Individual and group decision making under risk: An experimentalstudy of bayesian updating and violations of first-order stochastic dominace. *Journal of Risk and uncertainty*, *35*(2), 129–148.

Charness, G., & Levin, D. (2005). How psychological framing affects economic market prices in the lab and field. *The American Economic Review*, *95*(4), 1300–1309.

Chen, D. L., Schonger, M., & Wickens, C. (2016). otree - an open-source platform for laboratory, online and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Dekel, E., Lipman, B. L., & Rustichini, A. (1998). Standard state-space models preclude unawareness. *Econometrica*, *66*(1), 159–173.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on. *Perspectives on Psychological Science*, *6*, 274–290.

Dietrich, F. (2018). Savage's theorem under changing awareness. *Journal of Economic Theory*, *176*, 1–54.

Dominiak, A., & Tserenjigmid, G. (2021). Ambiguity under growing awareness. *Journal of Economic Theory*, 105256.

Eckel, C. C., & Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization*, *68*(1), 1–17.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, *10*(2), 171–178.

Gennaioli, N., & Shleifer, A. (2018). *A crisis of beliefs: Investor psychology and financial fragility.* Princeton University Press.

Gill, D., & Prowse, V. (2016). Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis. *Journal of Political Economy*, *124*(6), 1619-1676. doi: 10.1086/688849

Grant, S., Meneghel, I., & Tourky, R. (2017). Learning under unawareness. *Available at SSRN 3113983*.

Grant, S., & Quiggin, J. (2013). Inductive reasoning about unawareness. *Economic Theory*, *54*(3), 717–755.

Grant, S., & Quiggin, J. (2015). A preference model for choice subject to surprise. *Theory and Decision*, *79*(2), 167–180.

Greiner, B. (2015). An online recruitment system for economic experiments. *Journal of the Economic Science Association*, *1*(1), 114–125.

Grether, D. M. (1992). Testing bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior & Organization*, *17*(1), 31–57.

Halpern, J. Y., & Rêgo, L. C. (2008). Interactive unawareness revisited. *Games and Economic Behavior*, *62*(1), 232–262.

Heifetz, A., Meier, M., & Schipper, B. C. (2006). Interactive unawareness. *Journal of economic theory*, *130*(1), 78–94.

Hoffrage, R., Ulrich; Hertwig, & Gigerenzer, G. (2000). Hinsight bias: A by-product of knowledge. *Journal of Experimental Psychology*, *26*(3), 556-581.

Holt, A. M., Charles A.and Smith. (2009). An update on bayesian updating. *Journal of Economic Behavior & Organization*, *69*(2), 125–134.

Isoni, A., Graham, L., & Sugden, R. (2011). The willingness to pay—willingness to accept gap, the" endowment effect," subject misconceptions, and experimental procedures for eliciting valuations: Comment. *American Economic Review*, *101*(2), 991–1011.

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica*, *77*(2), 603–606.

Karni, E., Valenzuela-Stookey, Q., & Vierø, M.-L. (2020). Reverse bayesianism: A generalization. *The B.E. Journalof Theoretical Economics*, *forthcoming*.

Karni, E., & Vierø, M.-L. (2013). "Reverse Bayesianism": A choice-based theory of growing awareness. *American Economic Review*, *103*(7), 2790–2810.

Karni, E., & Vierø, M.-L. (2015). Probabilistic sophistication and reverse bayesianism. *Journal of Risk and Uncertainty*, *50*(3), 189–208.

Karni, E., & Vierø, M.-L. (2017). Awareness of unawareness: a theory of decision making in the face of ignorance. *Journal of Economic Theory*, *168*, 301–328.

Kochov, A. (2010). *A model of limited foresight* (Tech. Rep.). working paper, University of Rochester.

Modica, S., & Rustichini, A. (1999). Unawareness and partitional information structures. *Games and Economic behavior*, *27*(2), 265–298.

Ortoleva, P. (2012). Modeling the change of paradigm: Non-bayesian reactions to unexpected news. *American Economic Review*, *102*(6), 2410–36.

Piermont, E. (2021). Unforeseen evidence. *Journal of Economic Theory*(193), 105235.

Raven, J., Raven, J. C., & Court, J. (1998a). *Manual for raven's progressive matrices and vocabulary scales. section 4: The advanced progressive matrices.* Oxford, UK: Oxford Psychologists Press; San Antonio, TX: The Psychological Corporation.

Raven, J., Raven, J. C., & Court, J. (1998b). *Manual for raven's progressive matrices and vocabulary scales. section 1: General overview.* Oxford, UK: Oxford Psychologists Press; San Antonio, TX: The Psychological Corporation.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225–237.

Schipper, B. C. (2013). Awareness-dependent subjective expected utility. *International Journal of Game Theory*, *42*(3), 725–753.

Schipper, B. C. (2022). Predicting the unpredictable under subjective expected utility. *Mimeo*.

Shackle, G. L. S. (1949). *Expectation in economics.* Cambridge University Press.

Shackle, G. L. S. (1992). *Epistemics and economics: A critique of economic doctrines.* Transation Publishers.

Sonnemann, U., Camerer, C. F., Fox, C. R., & Langer, T. (2013). How psychological framing affects economic market prices in the lab and field. *PNAS*, *110*(29), 11779–11784.

Trautmann, S. T., & Schmidt, U. (2012). Pricing risk and ambiguity: The effect of perspective taking. *Quarterly Journal of Experimental Psychology*, *65*(1), 195–205.

Trautmann, S. T., & van de Kuilen, G. (2015). Ambiguity attitudes. In G. Keren & G. Wu (Eds.), *The wiley blackwell handbook of judgment and decision making* (pp. 89–116). Blackwall.

Trautmann, S. T., Vieider, F. M., & Wakker, P. P. (2011). Preference reversals for ambiguity aversion. *Management Science*, *57*(7), 1320–1333.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567.

Viscusi, W. K., & Chesson, H. (1999). Hopes and fears: the conflicting effects of risk ambiguity. *Theory and decision*, *47*(2), 157–184.

Chapter 2 Appendix

2.A Instructions for Experiment 1

Note: In red are comments, text highlighted in yellow relates to only the IS condition and text highlighted in green relates to only the PS condition

General Information

Welcome to the experiment.

Thank you for volunteering your time to participate in this experimental project. The purpose of this experiment is to study how people make decisions in a particular situation. The results of this experiment will have applications to behavioral economics and economics in general.

During this experiment, please follow the instructions very carefully. Please remain silent during the session. You will go through various stages where in some instances you will simply observe outcomes of draws from an urn and in other cases you will be asked to report your perceived likelihood of an event or how much an item is worth to you. These reports, which you will enter on your computer, will determine your eventual monetary reward from participating in this experiment.

You may have heard about experiments in which participants were deceived. This experiment ***does not*** *involve deception by the experimenters. That is, everything the experimenter tells you, and all on-screen instructions, are true and accurate.*

Initial Instructions

It is critical that you read through these instructions carefully as fully understanding them will allow you to substantially increase your eventual monetary payoff from this study, ***where you can earn from a minimum of €4.00 to a maximum of €26.00 depending on your decisions.***

During this study you will be asked to observe some outcomes of draws from an urn and will then be asked to report how likely certain 'events' are and how much some 'items' are worth to you. One of these choices will be randomly picked to determine your monetary payment for completing this study. Depending on your responses you stand to earn a substantial amount of money. Over the next few pages *we explain how your earnings will be determined. Please read this very carefully.*

Likelihoods of events – Reporting and Earnings

At different instances during this study, you will be asked to provide us with your perceived likelihood of an outcome of a random draw from an urn. The urn will contain prizes of varying monetary value. You will have an opportunity to observe multiple random draws out of this urn to gain a good understanding of the likelihood of different prizes.

After observing a sequence of random draws from the urn, you will be asked to report your beliefs about the likelihood that the prize drawn from the urn is of particular value. For example, we may ask you to report your perception of the likelihood that the prize drawn out of the urn has a value of €30.

You will be asked to report a number between 0 and 1. The closer to the true value your reported number is the greater would be your expected potential bonus from this decision.

Your best strategy is to estimate your own perceived likelihood and truthfully report that likelihood.

Suppose that we ask you to report the likelihood of drawing a ball with prize X. Your reported likelihood will be compared to a likelihood randomly generated by the computer. This ***randomly generated likelihood*** *will be a number between 0 and 1 and it will be completely unrelated to your reported likelihood. If your reported likelihood is greater than or the same as the randomly generated likelihood, you will be endowed with a lottery that pays you X with a probability equal to the actual proportion of prize X in the urn and pay you nothing otherwise. If, instead, your reported likelihood is lower than the randomly generated likelihood, you will be endowed with a lottery that pays you X with a probability equal to the randomly generated likelihood and pay you nothing otherwise. After these choices are made and revealed by the computer, the lottery in your (virtual) possession will be played and payments made according to the realization of the lottery.*

Example: Let's say we ask you to provide your perceived likelihood that a prize of value €30 will be randomly drawn from the urn. Let's further suppose that the true proportion of prize €30 is 0.5 in the urn. If your reported likelihood is 0.4 and the randomly generated likelihood is 0.3, you will receive the lottery that pays you €30 with probability 0.5 (true proportion of the prize) and pay you nothing otherwise. If your reported likelihood is 0.4 and the randomly generated likelihood is 0.6, you will receive the lottery that pays you €30 with probability 0.6 (the randomly generated likelihood) and pay you nothing otherwise.

Values of 'items' – Reporting and Earnings

At different instances during this study, you will be asked to provide us with the value at which you would be willing to sell an 'item'. This 'item' will be a lottery that would give you some monetary prizes with some probabilities. During the tasks that will follow, the prizes that will be possible to earn and their likelihoods will not be explicitly stated and so you would need to rely on what you expect.

For example, if you believe that there is equal likelihood of two prizes of value of €20 or €10, then the corresponding lottery would entail a payoff of €20 with probability 0.5 or a payoff of €10 with probability 0.5.

You will be given the lottery and it will be your task to provide us with a value which you would feel comfortable to sell us back this lottery. Your decision is essentially to provide us with the certain amount that you would be happy to receive instead of playing the lottery at the particular instance.

As you will see, your best strategy is to provide us with the minimum amount you would be willing to receive for selling us the lottery

Your named amount will be compared to a fixed amount. This fixed amount will be randomly generated by a computer and will be completely unrelated to your named amount:

- *If your named amount is less than or the same as the fixed amount, then you get to sell the lottery. But, here's the interesting part. You do not receive the amount you offered. Instead, we pay you the fixed amount, i.e. the randomly generated amount which is higher than or equal to your offer.*
- *If, on the contrary, your named amount is more than the fixed amount then you don't get to sell the lottery and will be paid according to the realization of the lottery.*

Example: if your named amount is €50 and the fixed amount is €60, you get to sell the lottery and receive the certain amount of €60. if your named amount is €50 and the fixed amount is €40, you do not get to sell the lottery and thus receive payment according to the realization of the lottery.

You should offer the minimum amount you would be willing to accept in exchange for the lottery you own. Your best strategy is to determine your personal value for the item and record that value as your offer. ***It will not be to your advantage to suggest more than this amount, and it will not be to your advantage to suggest less.*** *There is not necessarily a "correct" value. Personal values can differ from individual to individual.*

Example of best strategy for deciding valuation

The following example illustrates how you work out the minimum you are willing to accept for a lottery.

Imagine that I am a seller of a lottery "A". How do I know the minimum I'd be willing to sell lottery "A" for?

Start with 1 penny. Would I be willing to get 1 penny for the item? If NOT, then increase the amount to 2 pence. If I'm NOT willing to accept 2 pence, then increase further. I keep increasing until I come to an amount that makes me indifferent between keeping lottery "A" or getting a certain amount.

Example: Would I sell lottery "A" for €1.00? NO. So I need to consider higher amounts. Would I sell lottery "A" for €2.00? YES. Would I sell lottery "A" for €1.90? YES, Would I sell lottery "A" for €1.80, YES. Would I sell lottery "A" for €1.50? I don't care whether I end up with €1.50 or keep the lottery. Then that is the minimum I'd be willing to accept for lottery "A". I'll record that number on the computer.

The key to determining the minimum you'd be willing to accept is remembering that you will not necessarily *get* only *the amount you declare. Instead, if you receive anything, you will receive the fixed offer.*

Why is my best strategy to declare the minimum I'd be willing to accept? Let's go back to the example:

Say that I decide that the minimum I'd be willing to accept for lottery "A" is €1.50.

What happens if I declare more than €1.50? Say I declare €2.

If the fixed amount is, say, €1.90, then I don't sell the lottery. Had I declared €1.50, I would have received the amount €1.90 for a lottery that I think is worth €1.50. So I lose out.

What happens if I declare less that €1.50? Say I bid €1.00.

If the fixed offer is €1.20, then I have to accept €1.20 for a lottery that I really think is worth €1.50. I lose out.

Payment procedures

You will be asked to provide a value for lotteries and likelihood for events at different instances as we described in the previous pages. One out of all these decisions will be randomly chosen and payments will be made according to that decision and the realisation of the relevant lottery.

All prizes and lotteries that you will be asked to consider and make decisions about will be expressed in tokens. Each token corresponds to €0.05. Thus, a prize of 100 tokens will be equivalent to €5.

PART 1

You will now simply observe random draws of balls out of an urn. This urn contains a number of balls. Each ball represents a potential prize in terms of payment in tokens. You will observe 20 consecutive random draws with replacement from the urn. Please pay attention to the prizes that will appear and their frequencies. ***For IS conditions:*** *This urn contains* ***two and only two possible prizes.*** *Your earnings do not directly depend on the outcome of each random draw in this stage, but understanding the composition of the urn may considerably improve your future earnings.* ***For PS conditions:*** *At any point in the study new balls representing different tokens to what you have been observing so far may be added to this urn. Please click OK when you are ready to proceed.*

Belief Screen 1

We would now like to ask you some questions about the likelihoods of different prizes in the urn you have been observing.

Remember that your most profitable strategy is reporting truthfully your assessment of different likelihoods.

You can remind yourself of the payment procedure and instructions related to the task by referring back to the instructions in front of you.

What is your estimate of the likelihood that prize 80 is drawn from the urn? ______

What is your estimate of the likelihood that prize 190 is drawn from the urn? ______

WTA Screen 1

We would like to ask for your value of a lottery.

Recall that when answering questions about your value of a lottery, your named amount will be compared to a fixed amount. If your named amount is greater than or equal to the fixed amount then you do not sell the lottery and are thus paid according to the realization of the lottery. Otherwise, if your named amount is less than the fixed amount, you get to sell the lottery you have been endowed with and receive the fixed amount as a payment.

Remember that your most profitable strategy is reporting truthfully your valuations, you can remind yourself of the payment procedure and instructions related to the task by referring back to the instructions in front of you.

> *Thinking about the different prizes and the composition of the urn you have just observed.*
>
> *What is the minimum amount you are willing to accept to sell the lottery that pays according to a draw from the urn?* ______

After giving a choice, the following appears on the screen

Thank you. Your choice has been recorded.

For IS conditions:

We will now draw another prize from the urn.

If the decision you just made is selected by the computer to be the payment relevant round, the draw about to take place will be used to determine your payment.

You will not be shown the prize drawn out of the urn in this instance. Your colleague making the draws will make a note of the prize drawn to be used later if necessary.

We make a draw out of the original urn and the participant making the draws notes down the prize drawn.

At this point we bring the new urn, draw one prize and say:

On PC Screen:

*This urn contains **only** the prize you are about to be shown. Please click OK to confirm you understand this.*

Now we drop the contents of the original urn into the new urn which together form the updated urn.

For PS conditions:

If the decision you just made is selected by the computer to be the payment relevant round, the draw about to take place from the urn in the front will be used to determine your payment.

At this point we bring the new urn, draw one prize and say:

On PC Screen:

This urn contains new prizes. The urn contains no prizes similar to what you have been observing as a result of random draws from the other urn. Please click OK to confirm you understand this.

Now we drop the contents of the original urn into the new urn which together form the updated urn.

We will now draw a prize from the urn.

If the decision you just made is selected by the computer to be the payment relevant round, this draw will be used to determine your payment. You will not be shown the prize drawn out of the urn in this instance. Your colleague making the draws will make a note of the prize drawn to be used later if necessary.

A random draw from the updated urn is made and the participant making the draws notes down the prize drawn.

Belief Screen 2

We would now like to ask again some questions about the likelihoods of different prizes in the urn you have been observing.

Remember that your most profitable strategy is reporting truthfully your assessment of different likelihoods.

You can remind yourself of the payment procedure and instructions related to the task by referring back to the instructions in front of you.

What is your estimate of the likelihood that prize 80 is drawn from the urn? ______

What is your estimate of the likelihood that prize 190 is drawn from the urn? ______

What is your estimate of the likelihood that prize 15/375 is drawn from the urn? ______

WTA Screen 2

Remember that your most profitable strategy is reporting truthfully your valuations.

You can remind yourself of the payment procedure and instructions related to the task by referring back to the instructions in front of you.

Again thinking about the urn in front of you.

What is the minimum amount you are willing to accept to sell the lottery that pays according to a random draw from the urn? ______

> **After giving a choice, this appears:**
>
> *Thank you. Your choice has been recorded*
>
> *We will now draw another prize from the urn.*
>
> *If the decision you just made is selected by the computer to be the payment relevant round, this draw will be used to determine your payment. You will not be shown the prize drawn out of the urn in this instance. Your colleague making the draws will make a note of the prize drawn to be used later if necessary.*
>
> **A random draw from the updated urn is made and the participant making the draws notes down the prize drawn.**

PART 2

Lottery Choice Task

On your screen below you see a list of 6 lotteries with the prizes given in terms of tokens. You have to make a choice among these 6 lotteries. For each of the listed below lotteries, the chance for either of the two payoffs is equal. That is, for lottery 2 for example, you can win 24 tokens with 50% chance and 36 tokens with 50% chance. Your chosen lottery will be played out and you will be paid according to the realization of that lottery. As before, each token corresponds to €0.05 Thus, for a prize of 100 tokens the equivalent dollar amount will be €5.

Lottery	X	Y
1	28	28
2	24	36
3	20	44
4	16	52
5	12	60
6	2	70

PART 3

Short Raven Test implemented

PART 4

General Demographic Questionnaire

- How old are you? (years)
- What is your gender? (M/F/Other [Please describe if you wish]/Prefer not to disclose)
- What is your country of origin?
- What is your religion?
    - Budhist
    - Christian
    - Hindu
    - Jewish
    - Muslim
    - Sikh
    - No religion
    - Other [Please describe if you wish]
    - Prefer not to disclose
- What is your field of studies/major?
- What is your year of study?
- In high school, what was the highest possible grade? (E.g. A, 100, 20)
- What was your final grade in high school?
- In political matters, people talk of “the left” and “the right”. How would you place your views on this scale, generally speaking?

2.B Overview over Draws and WTAs in Experiment 1

Table A2.1: Sample draws per session & Average WTAs Reported.

Session	Treatment	Draw 1	Draw 2	Draw 3	Draw 4	Draw 5	Draw 6	Draw 7	Draw 8	Draw 9	Draw 10	Draw 11	Draw 12	Draw 13	Draw 14	Draw 15	Draw 16	Draw 17	Draw 18	Draw 19	Draw 20	WTA^o	WTA^u
1	IS, low prize	190	190	190	190	190	190	80	80	80	80	190	190	190	80	190	190	190	190	80	80	188.75	157.5
2	IS, low prize	80	80	190	190	190	190	190	190	190	80	80	190	190	190	190	190	190	80	80	80	94.875	72.625
3	IS, low prize	80	190	80	80	80	80	190	80	80	190	190	80	190	190	190	80	190	190	80	190	108.105	731.053
4	IS, low prize	80	190	190	80	190	80	190	80	190	80	190	190	190	190	190	190	190	190	190	190	119.364	756.364
5	IS, low prize	190	190	190	80	80	80	80	190	80	190	80	190	190	190	80	190	80	80	190	80	871.538	858.462
6	IS, low prize	190	80	190	80	190	80	190	80	80	190	190	190	190	190	80	190	80	190	80	190	134.429	125
7	IS, low prize	190	80	80	190	80	190	190	190	190	80	190	190	190	190	190	80	80	80	80	80	112.667	943.333
8	IS, high prize	80	80	80	190	190	80	190	80	190	190	190	190	190	190	190	190	80	80	190	80	115.333	164.667
9	IS, high prize	190	190	190	190	190	80	80	80	190	190	80	190	80	190	80	190	80	190	80	190	97.6	161
10	IS, high prize	190	190	190	190	190	190	190	80	190	80	190	80	190	190	190	80	80	80	190	190	752.143	974.286
11	IS, high prize	190	80	80	190	190	190	80	190	80	80	190	80	190	190	190	190	190	190	190	80	152.857	143
12	IS, high prize	190	80	190	80	190	190	190	190	190	190	190	80	80	190	80	190	80	80	190	190	117.714	185.643
13	IS, high prize	80	190	80	80	190	80	190	190	80	190	190	190	80	190	190	190	80	190	80	190	129.8	138
14	IS, high prize	190	80	190	80	190	190	80	190	80	80	80	80	190	80	190	190	190	80	80	80	110.778	171.111
15	PS, low prize	80	80	190	190	80	80	80	190	190	80	190	80	190	80	80	190	80	190	190	190	135	77.5
16	PS, low prize	190	190	190	190	190	190	80	80	190	80	190	80	80	190	80	80	80	80	80	190	124.25	96.05
17	PS, low prize	80	190	190	190	80	190	80	80	190	190	190	190	190	190	190	80	190	190	190	80	127.5	80
18	PS, low prize	190	190	190	80	190	190	80	80	190	80	80	190	190	190	80	80	190	190	80	190	115	766.667
19	PS, low prize	190	80	190	190	190	190	80	80	80	80	190	190	190	80	80	80	190	80	80	80	150	92.5
20	PS, low prize	190	190	190	190	80	190	190	80	190	190	190	190	190	190	190	190	190	80	190	190	154	125.333
21	PS, low prize	80	190	190	80	190	190	190	190	80	190	80	190	80	190	190	190	190	80	80	80	105.667	796.667
22	PS, low prize	80	190	190	80	80	190	80	80	190	80	190	190	190	190	80	190	80	190	190	80	179.286	960.714
23	PS, low prize	190	80	80	80	190	80	190	190	190	80	80	190	80	80	190	80	80	190	80	80	140.5	116
24	PS, low prize	80	80	190	190	190	80	190	80	190	80	190	190	80	190	80	80	80	80	190	190	127.273	823.636
25	PS, high prize	190	190	190	80	80	80	190	190	190	190	190	190	80	190	190	190	80	80	80	190	121.5	157.5
26	PS, high prize	190	80	80	190	190	190	190	190	80	190	80	80	190	80	190	80	80	190	80	190	131.944	173.889
27	PS, high prize	190	190	190	190	190	190	80	80	80	80	80	190	80	190	190	80	190	190	80	190	122.636	161.364
28	PS, high prize	190	190	190	80	190	80	80	80	190	190	80	80	190	190	190	190	190	80	190	80	124.308	174.462
29	PS, high prize	80	80	80	80	190	80	190	190	190	190	190	190	190	190	190	190	190	190	190	80	151.75	186.625
30	PS, high prize	190	190	190	80	190	190	190	80	190	190	190	190	190	80	80	80	190	80	190	190	149.7	180.1
31	PS, high prize	190	190	190	190	190	190	80	190	80	190	80	190	80	190	190	190	190	80	80	190	162.5	211.667
32	PS, high prize	190	80	190	80	190	80	190	80	80	190	190	190	80	80	190	190	80	190	190	190	134.375	175.625
33	PS, high prize	190	80	80	80	80	190	190	80	190	190	80	80	80	190	80	190	80	190	80	190	116.25	246.25

2.C Additional Analyses for Experiment 1

Figure A2.1: Histograms of the residual in the original urn ($\hat{p}^o_x = 1 - \hat{p}^o_{80} - \hat{p}^0_{190}$).

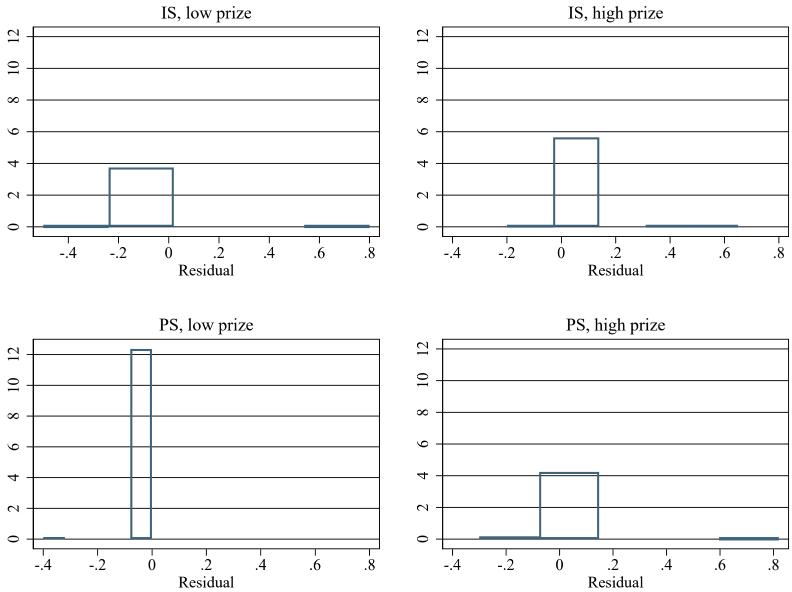


Table A2.2: Relation between the ratio differences and cognitive ability; OLS regression.

Ratio differences ΔR	
Cognitive ability	-0.008
	(0.006)
Constant	0.221*
	(0.099)
Observations	343

Notes: * $p < 0.05$; ** $p < 0.01$. Standard errors in parentheses. The estimation additionally controls for high prize, age, gender, being an economics student and risk aversion but the coefficients are not reported.

Figure A2.2: Histograms of the residual in the updated urn ($\hat{p}^u_x = 1 - \hat{p}^u_{80} - \hat{p}^u_{190} - \hat{p}^u_s$).

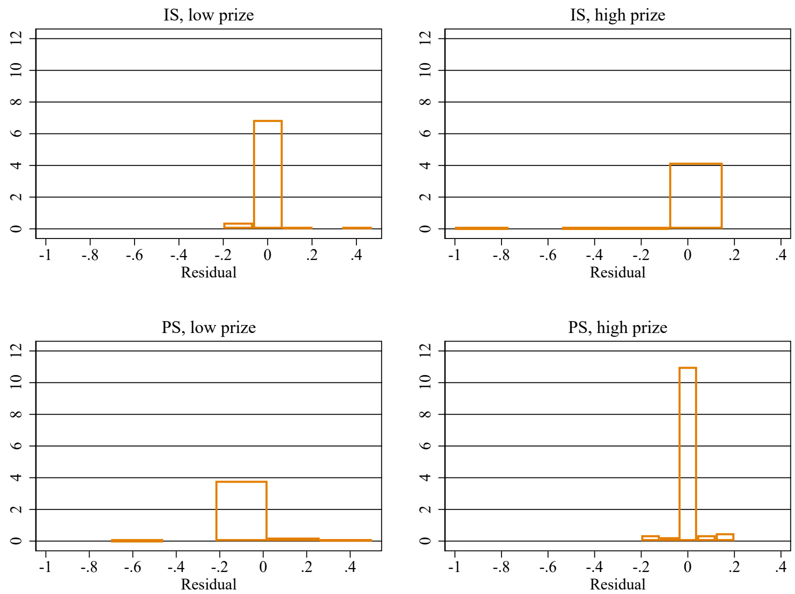


Figure A2.3: Histograms of the change in the ratios before vs. after the urn is updated, by treatment. Histogram in blue, box plot in orange, outliers (circles) and mean (diamond) in black.

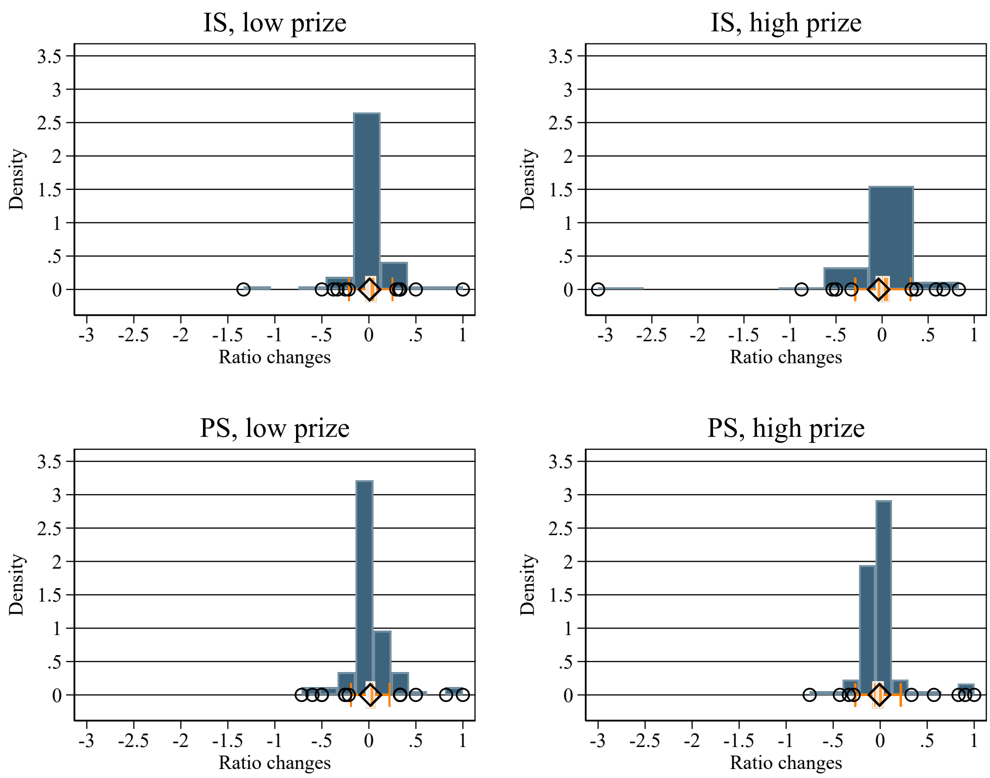


Table A2.3: Relation between *WTA* and possible moderators, original urn; OLS regression.

WTA	All treatments	Low prize	High prize
PS	23.770**	25.511*	23.797**
	(6.423)	(10.406)	(8.052)
# prizes 190 observed	-2.744	-1.238	-5.475
	(2.304)	(3.744)	(3.228)
Belief about # prizes 190	107.165**	90.308	133.747**
	(37.626)	(68.299)	(43.275)
Cog Ability	0.800	2.471	-1.416
	(1.527)	(2.619)	(1.872)
Age	-1.792*	-1.059	-2.869*
	(0.773)	(1.127)	(1.109)
Female	-20.848**	-24.145*	-18.878*
	(6.658)	(10.985)	(8.300)
Econ	1.172	4.272	-1.352
	(7.333)	(11.500)	(9.616)
Risk aversion	-1.200	0.967	-4.824
	(2.856)	(4.415)	(3.835)
Constant	126.481**	87.957*	186.385**
	(29.600)	(43.277)	(44.218)
Observations	344	169	175

Notes: * $p < 0.05$; ** $p < 0.01$.
Standard errors in parentheses.

Table A2.4: Relation between *WTA* and possible moderators, updated urn; OLS regression.

WTA	All treatments	Low prize	High prize
PS	16.509*	4.807	32.303**
	(7.722)	(6.924)	(10.866)
# prizes 190 observed	-0.420	-3.264	-6.300
	(2.404)	(2.046)	(3.868)
Belief about # prizes 190	80.578*	140.121**	-36.993
	(36.053)	(35.405)	(47.142)
Cog Ability	2.001	5.250**	0.365
	(1.838)	(1.767)	(2.486)
Age	-1.110	0.511	-1.665
	(0.930)	(0.756)	(1.461)
Female	-10.362	-5.076	-27.371*
	(7.983)	(7.374)	(11.097)
Econ	0.164	16.616*	-2.022
	(8.785)	(7.691)	(12.810)
Risk aversion	-7.285*	-3.280	-11.742*
	(3.400)	(2.879)	(5.139)
Constant	110.841**	14.018	304.474**
	(35.545)	(28.851)	(58.634)
Observations	344	169	175

Notes: * $p < 0.05$; ** $p < 0.01$.
Standard errors in parentheses.

2.D Instructions for Experiment 2

Note: In red are comments that were not visible to participants.

General Instructions

Thank you for participating in today's experiment.

If you have any questions during the experiment, please raise your hand. An experimenter will approach your table to answer your question in private.

You may have heard about experiments in which participants were deceived. This experiment does not involve deception by the experimenters. That is, everything the experimenter tells you, and all on-screen instructions, are true and accurate.

The experiment consists of 4 parts. For participating in this experiment you will earn £3 at the end of the experiment. In addition you can earn a bonus of £6 in each of the four parts, depending on your performance in the experiment and chance. After these four parts you will play a pattern game, in which you can earn additionally up to £2.

In the end follows a short demographic questionnaire.

Sampling Boxes

The experiment consists of 4 parts. In each part, you draw a random sample of (virtual) marbles from a (virtual) box containing exactly 100 colored marbles. Initially, you have no information about the contents of each box: you do not know which colors, or how many different colors, are in the box. The four parts and four boxes are independent of each other: different boxes are used for different parts.

In each part, you draw 30 marbles with replacement one after another from the box. You draw a marble by clicking the button "Draw" (or by pressing enter). Once clicked, the computer randomly draws a marble from the box. The result of a draw is shown on-screen with a marble of the color and the name of the color.

The sample draws are conducted with replacement. For example, if you drew a magenta marble (this color is not used in the actual experiment) from a box, this marble is placed back in the box for the next draw, such that the number of marbles of each color in the box stays the same as you sample. All marbles you have sampled (and their colors) are registered at the bottom of the screen.

Your payoff-relevant task

After each draw of a new marble, you will be asked to state your expectation about the contents of the box, that is, about the distribution of colors in the box. The more precise your prediction is, the higher will be your expected payoff from the experiment (details below). After each draw, you will be asked to separately indicate:

(i) Your expected number of marbles in the box for each color that you have already observed for the box, and

(ii) Your expected number of marbles of "any other colors" that you have not yet observed for the box, and that may or may not be in the box.

Example

Suppose you drew a magenta marble in your first draw and a teal marble (this color is also not used in the actual experiment) in the second draw. After the first draw you would be asked to guess how many magenta marbles are in the box, and how many marbles of any other color, not yet observed, are in the box. After the second draw you would be asked how many magenta marbles are in the box, how many teal marbles are in the box, and how many marbles of any other color, not yet observed, are in the box.

As the box contains exactly 100 marbles, your estimates of the number of marbles of the already observed colors and of any other colors you may think are in the box (but not yet observed) must add up to 100. Moreover, if you expect the number of marbles of other colors to be zero, you need to explicitly submit an estimate of zero (that is, not just leaving the entry field open).

After the 30$^{\text{th}}$ marble is drawn, you will enter your last prediction for this box. A new button "Continue to the next box" will allow you to continue to the next part, with a new box to sample.

Entering estimates in the program

After each draw, you can enter your estimates by typing them into the entry fields. You can also use the "fill previous estimate" buttons to pre-fill your previous round's estimates for each color. At any point before making the next draw, you can adjust the current estimates in the entry fields using "+" and "–" buttons next to the entry fields.

<u>Getting paid for good predictions</u>

You may earn a bonus of £6 for each part of the experiment. All of your answers provided for all four parts will affect your chances of receiving the bonus. If you want to maximize your expected earnings from this experiment, it is in your best interest to estimate the number of marbles for each box as accurately as possible, and report them truthfully after <u>each</u> draw. To determine whether you will win a bonus, you will draw a marble either from one of the boxes in the experiment (called *Estimate Box*), or from another, newly constructed one (called *New Box*). Importantly, your reported estimates will influence the construction of this new box.

If you report your estimates accurately and truthfully, this will be best for you in terms of your expected payment from the experiment. Below we will explain the <u>payment procedure</u>, and provide <u>the intuition</u> and an <u>example</u> why it is in your best interest to report your estimates as correctly as possible after each draw. You are invited to review these explanations. Please note that they are not necessary to understand the experiment and can be skipped without any harm if you are not interested. You can request a hard copy of these details at any point of the experiment in case of doubt.

Payment procedure (click to expand):

The below was hidden and only visible if the participants chose to expand the information:

After you finished sampling from all four boxes, for <u>each</u> of the four parts you may earn a bonus of £6 as follows:

Estimate box: The computer randomly selects one of the 30 draw rounds, and then randomly selects one color estimate you made for this round (this is the selected color for this task). This can be an estimate for some color you have observed, or alternatively an estimate for the number of not yet observed colors at some point, that is, “any other color”. Note than all of your estimates have the same chance to be randomly selected.

New box: Next, the computer constructs a new box of 100 marbles that contains only two colors, black or white. Every possible combination of black and white marbles (the number of white marbles = 100 – the number of black marbles) is equally likely.

Next, the computer compares the number of black marbles in the New Box with the estimate you made for the selected color in the experiment (or for “any other color”).

- If your estimate for the selected color is <u>larger</u> than the number of black marbles in the New Box, you will draw one marble from the <u>Estimate Box</u>. If this marble is of the selected color, you will receive £6. If the marble is not of the selected color, you will receive £0.

- If your estimate is smaller than the number of black marbles in the New Box, you will draw one marble from the New Box. If this marble is black, you will receive £6. If the marble is white, you will receive £0.

Intuition (click to expand):

The below was hidden and only visible if the participants chose to expand the information:

You will have the best chance to win the bonus of £6 for each part, by truthfully reporting your estimate. For example, if you think there are many magenta marbles in the Estimate Box, you will more likely make a draw from this box. This is because in your estimation the number of black marbles in the New Box will most likely be smaller than your estimate of magenta marbles for the Estimate Box.

If you think there are only few magenta marbles in the Estimate Box, you will more likely make a draw from the New Box. This is because the number of black marbles in the New Box will most likely be larger than your estimate of magenta marbles for the Estimate Box.

Thus, as long as you report your estimate for each color in each draw and each box accurately and truthfully, the mechanism makes sure that you get the box with the highest chance of winning the bonus.

Note that your winning chance in the case of making the payoff-relevant draw from the Estimate Box depends only on the true number of marbles of that color in the box. Similarly, in the case of making the payoff-relevant draw from New Box, the chance depends only on the number of black marbles in the box. Your estimate of colors for the boxes in the experiment is only relevant for determining the best boxes for you during the payment procedure. Thus, better estimates give you better chances to win.

Example (click to expand):

The below was hidden and only visible if the participants chose to expand the information:

Example - Part 1

For part 1, the computer selected the round 16 draw. In this round you provided estimates of the number of magenta marbles, teal marbles, and the number of marbles of “any other color”.

The computer further selected magenta as the payoff-relevant color estimate. Suppose your estimate of the number of magenta marbles in box 1 in round 16 was 42 marbles.

Suppose the computer randomly generated a New Box that contained 35 black and 65 white marbles. Because 35 black winning marbles in New Box is less than your estimate of 42 magenta winning marbles in Estimate Box 1, your bonus would be determined by Estimate Box 1. Note that your true chance to win the bonus of £6 would depend on the true number of magenta marbles in box 1. Suppose you drew a teal marble from Estimate Box 1. Your bonus for part 1 would be £0.

Example - Part 2

For part 2 box, the computer selected the round 2 draw. In this round you provided estimates of the number of magenta marbles, and the number of marbles of “any other color”. The computer further selected “any other color” as the payoff-relevant color estimate. Suppose your estimate for the number of “any other color” marbles in box 2 in round 2 was 50 marbles.

Suppose the computer randomly generated another New Box that contained 7 black and 93 white marbles. Because 7 black winning marbles in New Box is less than 50 winning marbles of “any other color” in Estimate Box 2, your bonus would be determined by a draw from Estimate Box 2. Note that your true chance to win the bonus of £6would depend on the true number of - marbles in box 2 that are not magenta. Suppose you drew a teal marble from box 2. Your bonus for part 2 would be £6.

Example - Part 3

For part 3 box, the computer selected the round 30 draw. In this round you provided estimates for the number of magenta marbles, the number of teal marbles, and the number of marbles of “any other color”. The computer further selected teal as the payoff-relevant color estimate. Suppose your estimate for the number of teal marbles in box 3 in round 30 was 20 marbles.

Suppose the computer randomly generated another New Box that contains 87 black and 13 white marbles. Because 87 black winning marbles in New Box is more than 20 twinning marbles of teal color in Estimate Box 3, your bonus would be determined by New Box 3. Suppose you drew a black marble from box 3. Your bonus for part 3 would be £6.

Example - Part 4

For part 4 box, the computer selected the round 7 draw. In this round you provided estimates for the number of magenta marbles, the number of teal marbles, and the number of marbles of “any other color”. The computer further selected “any other color” as the payoff-relevant color estimate. Suppose your estimate for the number of “any other color” marbles in box 4 in round 7 was 33 marbles.

The computer randomly generated another New Box that contains 27 black and 73 white marbles. Because 27 black winning marbles in New Box 4 is less than 33 winning marbles of “any other color” in Estimate Box 4, your bonus would be determined by a random draw from Estimate Box 4. Note that your true chance to win the bonus of £6would depend on the true number of marbles in box 4 that were neither magenta nor teal. Suppose you drew a teal marble from box 4. Your bonus for part 4 would be £0.

Pattern game

You will now play a pattern game, where you are asked to solve some puzzles

On the screen, you will see a set of abstract pictures with one of the pictures missing. You need to choose a picture from the choices below to complete the pattern.

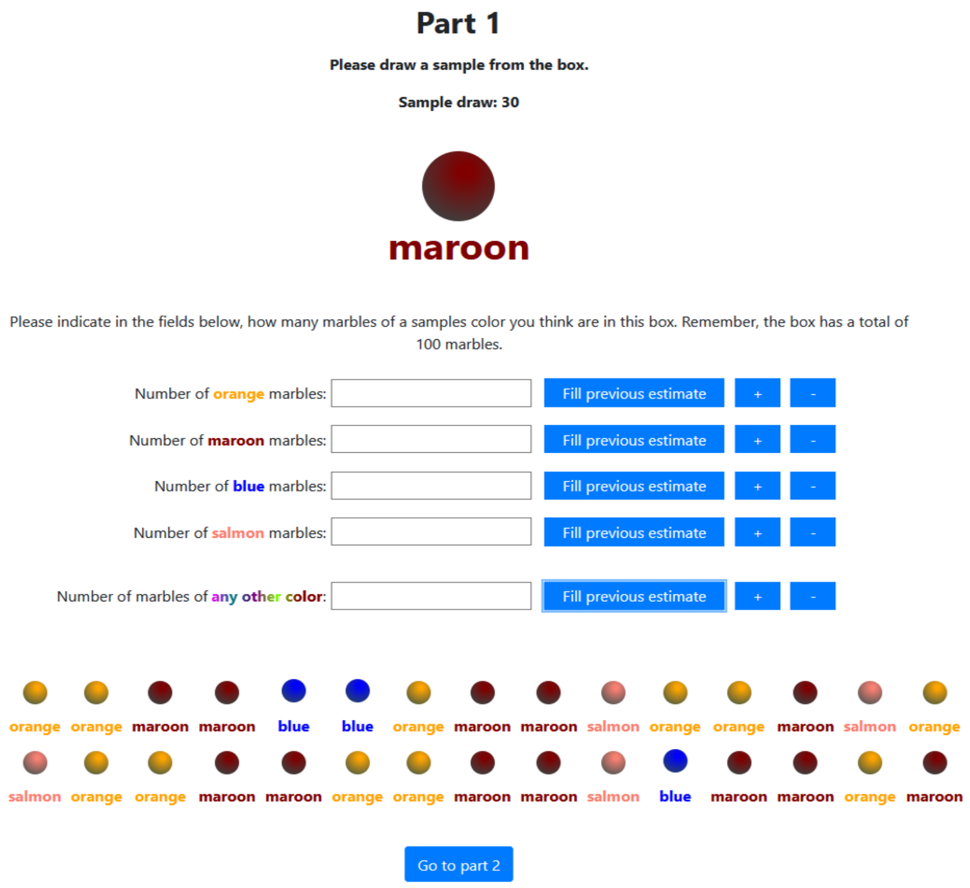
You will have a total of 8 minutes to complete 12 such puzzles.

During these 8 minutes you will be able to move forwards and backwards and change your answers using the buttons and tabs on your screen.

At the end of the experiment, the computer will randomly draw two of the puzzles from the pattern game. Each puzzle has the same probability to be chosen. For each of the two puzzles that you solved correctly, you will earn an additional £1.

Once the 8 minutes have passed, the pattern game will be automatically submitted and you will proceed to the results. You can submit all your answers and wait for the others to finish once you reach the last puzzle by clicking on the button that will appear and be labelled "Finish and go to results".

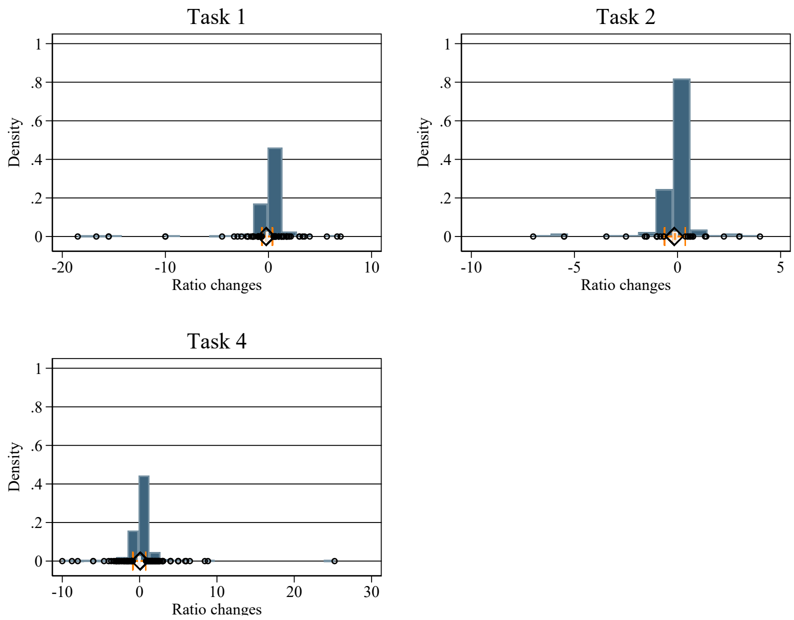
2.E Sample Screens for Experiment 2



Notes: Example of the sampling screen in the experiment, first task, four outcomes first treatment. The example depicts the screen after the 30th sample draw.

2.F Additional Analyses for Experiment 2

Figure A2.4: Histograms of the residual $\hat{p}_x^{30}$ (after observing the last sample draw).



Notes: The orange boxes show the residuals of *two colors* treatment while the the blue boxes show the residuals of *four colors* treatment.

Table A2.5: Relation between the ratio differences and cognitive ability, panel GLS.

Ratio differences ΔR	
Cognitive ability	0.028
	(0.026)
Num. draws	0.022*
	(0.010)
Num. colours observed	-0.066
	(0.123)
Constant	-0.314
	(0.651)
Observations	1,119
Subjects	

Notes: * $p < 0.05$; ** $p < 0.01$. Standard errors in parentheses. The estimation additionally controls for age, gender, being an economics student and risk aversion but the coefficients are not reported.

Figure A2.5: Histograms of the changes in the ratios before vs. after the urn is updated, by treatment. Histogram in blue, box plot in orange, outliers (circles) and mean (diamond) in black.

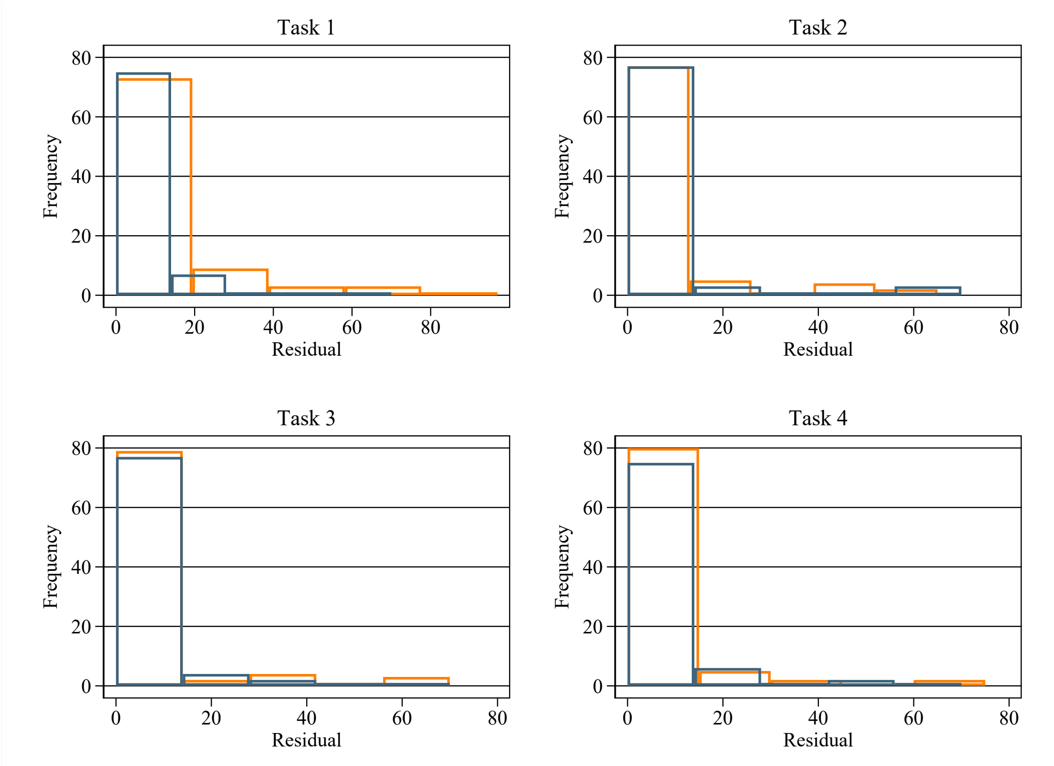


Table A2.6: Differences in the ratio changes, depending on whether participants expected a surprise ($\hat{p}^t_x$>0).

		Didn't expect surprise	Expected surprise	p-value	p-value (corr)
Task 1	ΔR^3	5	80	0.664	1.000
	ΔR^4_1	5	79	0.919	1.000
	ΔR^4_2	5	79	0.202	1.000
	ΔR^4_3	5	79	0.466	1.000
Pooled	ΔR^4_P	15	237	0.221	1.000
Task 2	ΔR^3	54	115	0.698	1.000
Task 4	ΔR^3	38	135	0.096	1.000
	ΔR^4_1	33	131	0.330	1.000
	ΔR^4_2	33	130	0.430	1.000
	ΔR^4_3	33	130	0.343	1.000
Pooled	ΔR^4_P	99	391	0.134	1.000

Notes: Wilcoxon rank-sum test, p-values corrected by Bonferroni-Holm procedure.

Table A2.7: Regression: relation between the number of sampled marbles and the residuals, panel fixed effects.

Ratio differences ΔR	
Num. draws	-0.735**
	(0.046)
Num. colours observed	-8.495**
	(0.532)
Constant	52.451**
	(1.868)
Observations	20,880
Subjects	174

Notes: * $p < 0.05$; ** $p < 0.01$.
Standard errors in parentheses, standard errors are clustered at the individual level.

Table A2.8: Regression: relation between the number of sampled marbles and the residuals after sample round 15, panel GLS.

Ratio differences ΔR	
Num. draws	-0.349**
	(0.051)
Num. colours observed	0.852
	(0.795)
Cognitive ability	-2.299**
	(0.517)
Four colours first	-2.800
	(2.250)
Constant	35.562**
	(11.238)
Observations	9,900
Subjects	165

Notes: * $p < 0.05$; ** $p < 0.01$.
Standard errors in parentheses, standard errors are clustered at the individual level. The estimation additionally controls for age, gender, being an economics student and risk aversion but the coefficients are not reported.

Figure A2.6: Residuals and estimates of a new color before and after observing a new color, second color.

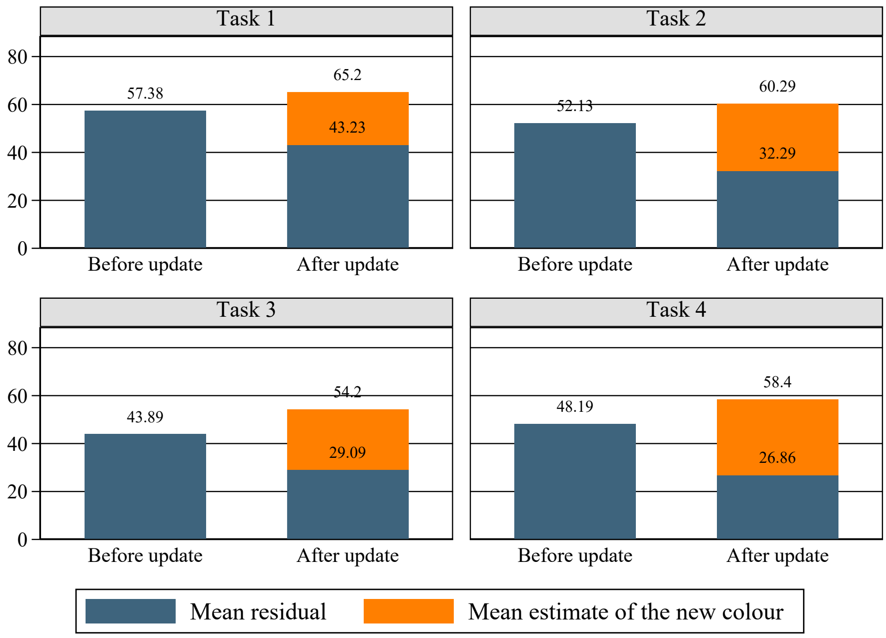
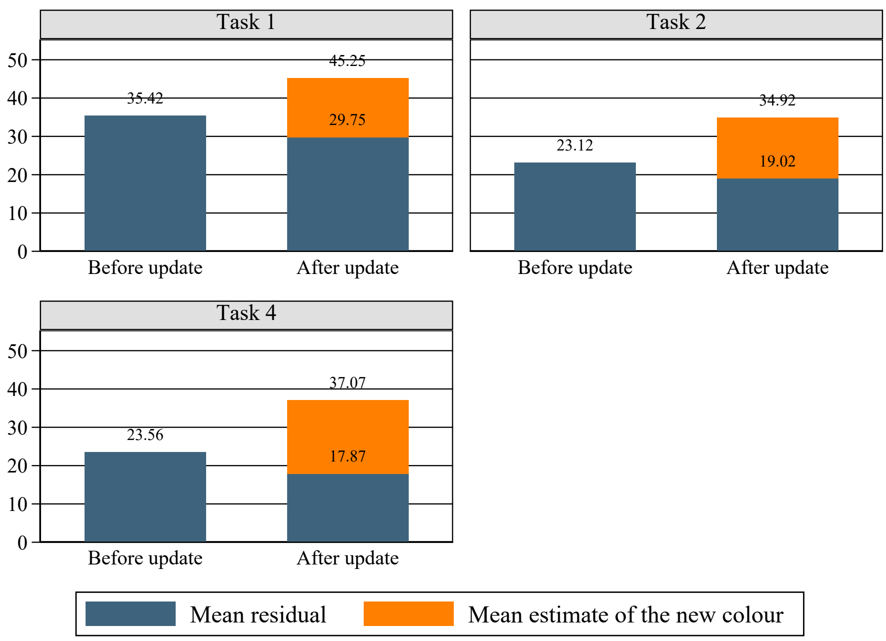


Figure A2.7: Residuals and estimates of a new color before and after observing a new color, third color.



Chapter 3

Measuring Inflation Expectations: How the Response Scale Shapes Density Forecasts

Abstract[§]

In density forecasts, respondents are asked to assign probabilities to pre-specified ranges of inflation. In two large-scale experiments, one conducted in the US and one in Germany, we show how answers vary when we modify the response scale: Shifting, compressing or expanding the scale leads to shifted, compressed and expanded forecasts. Mean forecast, uncertainty, and disagreement vary by several percentage points. The results have implications for survey design and for central banks' optimal adjustment of the response scales during times of high inflation.

[§]Joint work with Peter Duersch and Thomas Eife.

We thank Kenza Benhima, Muhammed Bulutay, Christian Conrad, Timo Dimitriadis, Zeno Enders, Frank Heinemann, Luba Petersen, Shyam Sunder, and Michael Weber for their helpful comments. We are especially grateful to Norbert Schwarz for his advice and for answering our many questions. We also would like to thank participants of the following conferences and seminars: LAGV in Marseille, HKMetrics in Mannheim, EF in Bonn, WTEM and DIW in Berlin. Financial support from the University of Heidelberg is gratefully acknowledged. Finally, we would like to thank the Bundesbank for including our questions in its survey on consumer expectations.

3.1 Introduction

Managing inflation expectations is an important part of modern central banking. When interest rates reached levels around zero after the financial crisis of 2008, central banks widely adopted this non-conventional policy tool. Managing expectations requires measuring expectations, and several new surveys have been established in the past decade, many using density questions. In this question format, respondents are given a response scale with pre-specified intervals and are asked to assign probabilities to the intervals that best represent their beliefs about inflation.

The experimental results we present in this paper show how the specifics of the response scale determine the responses. Shifting or compressing the response scale causes respondents to shift or compress their answers. For example, we can vary respondents' mean inflation forecast from -0.32% to 8.15% simply by shifting the response scale. Similarly, we can double respondents' average uncertainty (the standard deviation of their response) from 3.08% to 6.08%, when we double the width of the scale. While these examples are extreme, it is clear that density forecasts cannot provide information about how well respondents' inflation expectations are "anchored" around a certain value (e.g., around the central bank's target) if the scale is not taken into account. Differencing inflation beliefs to obtain changes of expectations over time does not solve the problem since the distortion itself can change over time.

Survey researchers have long been aware that even minor variations in the wording of a question or in the design of a questionnaire may strongly affect the responses (see Schwarz, 2010, for an overview and Payne, 1951, and Sudman & Bradburn, 1974, for early contributions). The recent literature on inflation expectations also addresses these points. Phillot & Rosenblatt-Wisch (2018) discuss the effect of question ordering on respondents' forecast consistency. The effect of a question's wording is addressed in several papers. Bruine de Bruin et al. (2012) study whether asking for "prices in general", "inflation", or "prices you pay" affects the responses and conclude that inflation expectations were lower and less dispersed when asking for "inflation" (see also Manski, 2018; Bruine de Bruin et al., 2023). Asking for the "overall inflation rate" or for "prices overall in economy" does not appear to systematically affect the results (Coibion et al., 2020). Providing additional information in the question (e.g., a newspaper article or a statement about the Federal Reserve's inflation target) affects households' responses (Coibion et al., 2022).

Our focus here is on variations of the response scale. In an influential study, Schwarz et al. (1985) show that shifting the response scale in an interval question (where respondents are asked to pick a single interval) may shift the responses. This phenomenon is a robust finding in survey research and has been replicated in various other studies (Schwarz, 2010, gives an overview). We extend this research agenda to density questions. Using the New York Fed's Survey of Consumer Expectations (SCE) question on inflation expectations as our baseline, we employ a battery of 12 treatments to systematically test whether and how changes to the scale affect the results. Four treatments study the effect of shifting the response scale and four treatments study the effect of compressing or expanding the scale. The final four treatments study the consequences of the irregular spacing of the response scale of the SCE, where the four center intervals are narrower than the other closed intervals. That is, the final four treatments study the effect of combining or splitting up existing intervals.

We collect data on two different subject pools: A representative sample of 1,300 respondents from the United States, on which we ran all 13 treatments, and a representative sample of more than 4,000 respondents from Germany on which the Bundesbank (Germany's central bank) ran three of our treatments.

The rest of the paper is organized as follows. Section 3.2 describes the experimental design and provides details on the hypotheses. Section 3.3 presents the results of the US survey and Section 3.4 the results of the German survey. Section 3.5 interprets the results, discusses possible improvements of density questions and how central banks can adjust the response scales during times of high inflation. Section 3.6 concludes.

3.2 Experimental design

We use three questions from the New York Fed Survey of Consumer Expectations (SCE) in our experiment. First, survey respondents are asked to provide a density forecast for 12-month ahead inflation (question Q9 in the SCE). Second, respondents report whether they expect inflation or deflation in a binary question (Q8v2 in the SCE). Third, we ask for a point forecast (Q8v2part2 in the SCE). Since we are primarly interested in the density forecast, our ordering of the questions differs from the ordering of the SCE. We move the density forecast in first place to prevent the other two questions (especially the point forecast) from confounding the responses of the density forecast. Several other surveys, such as the Bundesbank household survey, adopted the response scale of the SCE.[26]

We use the response scale of the SCE as our *Baseline* treatment. The scale has ten intervals, including two open outer intervals, and is centered at zero. The closed intervals range from −12% to 12%. The four central intervals are narrower than the outer closed intervals, see Figure 3.1. As in the SCE, the question asking for a point forecast varies depending on whether respondents expect inflation or deflation in the binary question. When a respondent expects deflation in the binary question, the point forecast asks for a deflation rate. When a respondent expects inflation, the point forecast asks for an inflation rate.

After each of these three questions, respondents are asked to indicate how certain they feel about their answer on a 6-item Likert scale (ranging from Very Uncertain to Very Certain). Following these six main questions, respondents are asked to answer a short questionnaire about their age, gender, political orientation and state of residence. Additionally, the questionnaire includes three measures of potentially relevant knowledge: A question on highest education degree obtained, a question on their knowledge of the Fed's inflation target, and three questions on financial literacy taken from Lusardi & Mitchell (2014). Finally, the questionnaire includes

[26] Bruine de Bruin et al. (2023) discuss the history of eliciting expectations in economics and provide an overview of current surveys that use probabilistic (density) questions.

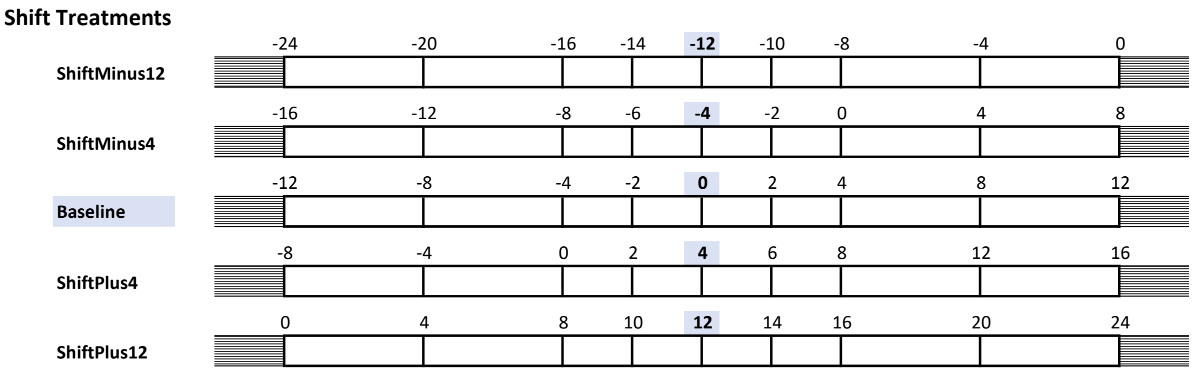
a control question to test the attentiveness of the respondents.

In the survey, respondents face one of 13 different treatment conditions. The response scale in *Baseline* is identical to the SCE. The other 12 treatment conditions introduce different variations to the response scale. All other questions are the same in all treatments. Hence, our design allows us to isolate the effect of changing the scale of the density forecast on the forecast itself, but also on subsequent assessments of 12-month ahead inflation via other question types. The 12 treatment conditions are grouped into three categories: *Shift* treatments, *Compression* treatments and *Centralization* treatments. The following three subsections present the different categories in greater detail.

3.2.1 Shift treatments

In the *Shift* treatments, the response scale is shifted towards either inflation or deflation, keeping all other parameters (number of intervals and their relative widths) constant. This means that the center of the scale moves away from zero compared to *Baseline*. The *Shift* treatments allow us to test how respondents' forecasts are influenced by different positions of the scale on the number line. We implement both shifts in two different degrees, resulting in a total of four *Shift* treatments: *ShiftMinus12*, *ShiftMinus4*, *ShiftPlus4*, and *ShiftPlus12*. Figure 3.1 illustrates the four *Shift* treatments, with *Baseline* as a reference. In *ShiftMinus12* and *ShiftMinus4* we subtract 12 and 4 respectively from all interval limits. Conversely, in *ShiftPlus4* and *ShiftPlus12*, we add 4 and 12 to the interval limits.

Figure 3.1: Response scales used in the **Shift treatments** (with Baseline for reference).

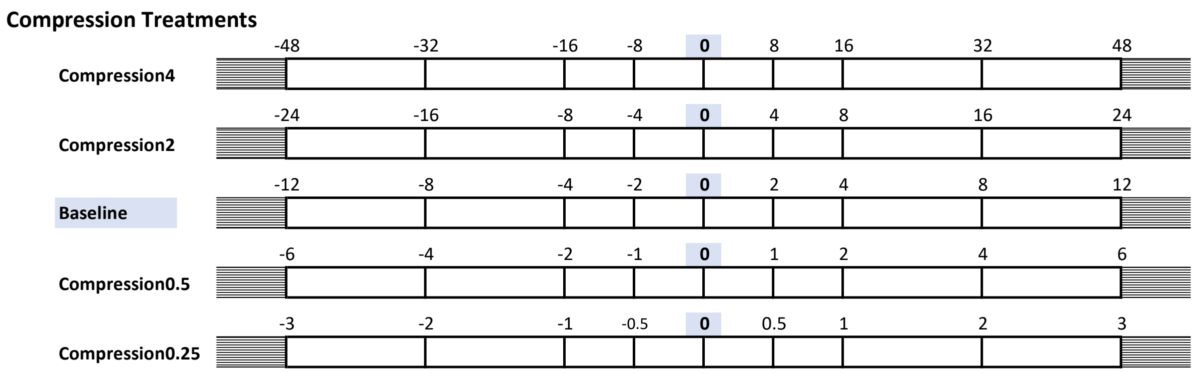


3.2.2 Compression treatments

In the four *Compression* treatments, the interval limits of the response scale are multiplied by a constant factor, keeping the number and the relative size of the intervals unchanged. For factors below 1, this leads to a compression of the response scale around the center. Factors above 1 result in an expansion (decompression). As before, we implement both compression and decompression with two different degrees, giving us four *Compression* treatments: *Compression0.25*, *Compression0.5*, *Compression2*, and *Compression4*.

In *Compression0.25* and *Compression0.5* the interval limits are multiplied by 0.25 and 0.5 and thus provide scales that zoom in more closely to inflation rates close to zero. In contrast, *Compression2* and *Compression4* widen the intervals. As Figure 3.2 illustrates, this results in values now being explicitly included in intervals that would have been part of the open intervals in *Baseline*. While *Compression2* and *Compression4* thus allow respondents to better communicate beliefs about high inflation and high deflation rates, they also imply a coarser image of respondents' inflation beliefs around the center.

Figure 3.2: Response scales used in the **Compression treatments** (with Baseline for reference).



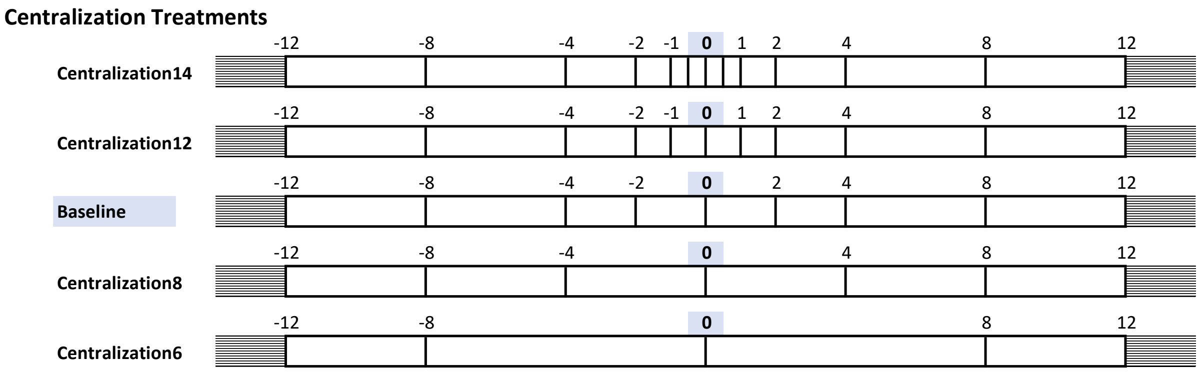
3.2.3 Centralization treatments

Finally, the four *Centralization* treatments vary the number of intervals around the center of the scale. Differently to the other two treatment categories, where the scale is either shifted or compressed, the overall span of the scale is identical to *Baseline*. Instead, we either split existing intervals around the center or we combine them, thus changing the number of intervals. Similarly to the *Compression* treatments, this allows for a finer or coarser image of respondents' inflation beliefs around the center, however, without changing the span of the scale itself. As with the other treatment categories, the centralizations are implemented with two different degrees, giving us the final set of four treatments: *Centralization6*, *Centralization8*, *Centralization12*, and *Centralization14*. Figure 3.3 depicts all four treatments relative to *Baseline*. In *Centralization6* and *Centralization8* the center intervals are combined, such that the overall number of intervals decreases to 6 or 8, respectively. In *Centralization8* all closed intervals have the same width. Respondents in these treatments thus can only state very coarse beliefs. In *Centralization12* and *Centralization14*, on the other hand, we split the intervals around the center allowing respondents to more finely express beliefs in this range.

3.2.4 Hypotheses

The design of the experiment and the hypotheses were pre-registered on the AEA RCT registry (`www.socialscienceregistry.org/trials/8716`). The study received ethics approval from the German Association for Experimental Economic Research (`https://gfew.de/ethik/`

Figure 3.3: Response scales used in the **Centralization treatments** (with Baseline for reference).



apyKIJdX).

Across treatment hypotheses

Regarding the *Shift* treatments, if respondents use the scale as a reference, we would expect the responses to shift in the same direction.

Hypothesis 1: *In the Shift treatments, the reported distributions of inflation expectations shift in the direction of the scale shift.*

After the density forecast, respondents answer the binary inflation/deflation question and the point forecast. If the treatment interventions from the density forecast carry over to the two subsequent inflation questions, we expect the responses in the *Shift* treatments to differ from *Baseline*. The *Shift* treatments provide two intuitive predictions to test:

Hypothesis 2: *In the Shift treatments, the incidence of expecting deflation is lower [higher] for positive [negative] shifts of the scale. The incidence of expecting inflation is higher [lower] for positive [negative] shifts of the scale.*

Hypothesis 3: *In the Shift treatments, the point forecast is higher [lower] for positive [negative] shifts of the scale.*

In the *Compression* treatments we compress or expand the entire scale. If respondents use the scale as a reference, they should compress or expand their belief distribution. Thus, we expect the dispersion to differ from *Baseline*.

Hypothesis 4: *In the Compression treatments, the reported distributions are more [less] dispersed in the less [more] compressed treatments.*

The *Centralization* treatments split or merge the intervals around the center of the scale. After splitting an interval into two smaller intervals, respondents can still provide the same response, but earlier literature has demonstrated that splitting and merging affects the responses. The sum of probabilities assigned to a subset of events typically exceeds the probability assigned to the overarching event (Tversky & Koehler, 1994; Sonnemann et al., 2013). Following this logic, splitting intervals around the center of the scale would lead to more probability mass being concentrated around the center. Accordingly, we would expect the dispersion being affected by the number of intervals around the center.

Hypothesis 5: *In the Centralization treatments, the reported distributions are more [less] dispersed, if the number of intervals in the central part of the scale is lower [higher].*

Within treatment hypotheses

In the within-treatment hypotheses, we study respondents' internal consistency and how our results are moderated by personal characteristics. We define consistency as a respondent's point forecast being compatible with the respondent's density forecast.

Hypothesis 6: *Subjects report consistent inflation forecasts.*

The effects of our treatment interventions might depend on a respondents' *proficiency* concerning monetary policy. As outlined above, we use three measures to capture different aspects of a respondents' proficiency. Respondents with a higher financial literacy or higher education level might be better informed about monetary policy and thus be less susceptible to changes to the scale. Similarly, respondents that know the inflation target of the central bank might be more anchored towards this target, expecting the central bank to rein in the inflation rate if it deviates from the target. Coibion et al. (2018), for example, show that managers' inflation expectations strongly react to receiving information about the central bank's inflation target. Additionally, such respondents might also feel surer that their answers are correct. In line with these deliberations, we test two further, directional hypotheses:

Hypothesis 7: *Respondents with better education/financial literacy/knowledge of the inflation target are affected less by the treatment interventions.*

Hypothesis 8: *Respondents with better education/financial literacy/knowledge of the inflation target are more certain in their answers.*

3.3 The US survey

We conducted the survey in the US in December 2021. For this month, the Bureau of Labor Statistics reports year-on-year inflation of 7.2 percent. This is somewhat higher than in the preceding five months, where inflation averaged at around 6 percent. Especially energy prices had been increasing in the months before the survey. For November 2021, for example, the Bureau of Labor Statistics reports a year-on-year increase in the price of energy of more than 50 percent.

3.3.1 Implementation

The US survey used all 13 treatments, was programmed in oTree (Chen et al., 2016), and was conducted on Prolific (`www.prolific.co`), a UK-based commercial subject pool.[27] On Prolific,

[27]See Appendix 3.A for the instructions and Appendix 3.B for screenshots of the three inflation questions.

Table 3.1: Descriptive statistics by treatment.

Treatment	**Response scale**			**Demographics**				
	#	Center	Span	Obs	Avg. age	Share female	Share white	Share black
Baseline	10	0	24	101	45.25	0.48	0.71	0.13
ShiftMinus12	10	-12	24	99	44.45	0.45	0.74	0.11
ShiftMinus4	10	-4	24	99	47.09	0.41	0.78	0.12
ShiftPlus4	10	4	24	98	43.46	0.47	0.68	0.18
ShiftPlus12	10	12	24	98	43.64	0.48	0.77	0.15
Compression4	10	0	96	99	46.75	0.61	0.84	0.05
Compression2	10	0	96	99	43.70	0.51	0.72	0.15
Compression0.5	10	0	12	96	45.09	0.47	0.74	0.18
Compression0.25	10	0	6	100	45.80	0.61	0.82	0.09
Centralization14	14	0	24	96	44.90	0.45	0.73	0.15
Centralization12	12	0	24	96	43.85	0.48	0.71	0.14
Centralization8	8	0	24	99	46.08	0.56	0.85	0.11
Centralization6	6	0	24	99	44.18	0.46	0.75	0.14
Average				**98.4**	**44.95**	**0.49**	**0.76**	**0.13**

Notes: Number of intervals (#), center of response scale, span of the closed intervals, number of respondents (obs), average reported age, and percentage share of female. The last two columns show the share of people identifying as white or black as recorded by Prolific.

we recruited a representative sample of the US population (stratified along sex, age and race). Data collection started on December 17th and finished on December 19th 2021.

In total, 1301 respondents completed our survey, with 100 respondents per treatment condition, except *Baseline*, which had 101 respondents. For the data analysis, we dropped 22 respondents: One failed the attention check, one appeared to reside outside the US and 20 provided beliefs in the density forecast that did not add up to 100 (see Table 3.1).[28] Respondents were paid a fixed amount of 1 (worth $1.33 at the time of the experiment) for completing the survey. On average, it took respondents 5:44 minutes to finish the survey. Based on our payment, respondents earned on average an hourly wage of $16.40, well above the average hourly earnings on Prolific.

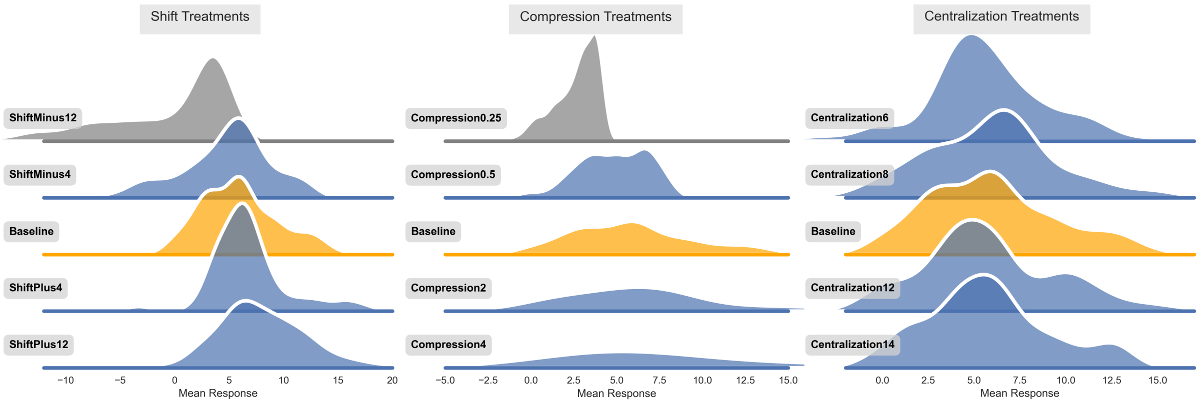
3.3.2 Results

Does the response scale affect the survey responses? Figure 3.4 gives a first impression. The figure shows the distribution of respondents' mean inflation expectations for all treatments. The *Shift* treatments are shown on the left, the *Compression* treatments in the center, and the *Centralization* treatments on the right. Moving the intervals of the scale to the left or right in *Shift* moves responses in the same direction. Similarly, compressing or expanding the scale in *Compression* also compresses and expands the answers.

Table 3.2 shows, for each of the 13 treatments, the average mean forecast, the average

[28]Respondents whose probabilities do not add up to 100 are prompted once to correct their answer. However, submitting an answer whose probabilities do not sum up to 100 was possible. For more detailed attrition and randomization checks, see Appendix 3.D.

Figure 3.4: Distribution of mean forecasts.



Notes: Kernel density estimates by treatment. *Shift* treatments in the left panel, *Compression* treatments in the center panel, and *Centralization* treatments in the right panel. Each panel uses a common y-axis with *Baseline* shown in orange in the center for comparison. Treatments with large probability mass in the open intervals in gray. Mean forecasts are calculated using a mass-at-midpoint assumption.

Table 3.2: Treatment differences for the US survey.

Treatment	**Statistics**									
	Mean Forecast				**Forecast Uncertainty**				**Disagree-ment**	
	beta		**m.a.m.**		**beta**		**m.a.m.**		**beta**	**m.a.m.**
Name	Avg.	P-value	Avg.	P-value	Avg.	P-value	Avg.	P-value		
Baseline	5.56		5.87		3.08		3.81		3.48	3.63
ShiftMinus12	-0.32	(1) 0.000 ***	-0.53	(1) 0.000 ***	3.78	(2) 0.037 **	3.36	(2) 0.228	5.87	5.67
ShiftMinus4	4.31	(1) 0.011 **	4.64	(1) 0.014 **	3.30	(2) 0.438	4.10	(2) 0.333	4.13	4.26
ShiftPlus4	6.59	(1) 0.019 **	6.83	(1) 0.030 **	3.38	(2) 0.347	4.15	(2) 0.308	3.51	3.58
ShiftPlus12	8.15	(1) 0.000 ***	8.34	(1) 0.000 ***	3.54	(2) 0.035 **	4.07	(2) 0.181	4.42	4.46
Compression4	10.98	(2) 0.000 ***	11.77	(2) 0.000 ***	8.85	(1) 0.000 ***	11.09	(1) 0.000 ***	11.55	12.39
Compression2	6.23	(2) 0.353	6.76	(2) 0.237	6.08	(1) 0.000 ***	7.61	(1) 0.000 ***	6.34	6.54
Compression0.5	4.55	(2) 0.013 **	4.81	(2) 0.012 **	1.84	(1) 0.000 ***	2.16	(1) 0.000 ***	2.00	2.03
Compression0.25	2.61	(2) 0.000 ***	2.66	(2) 0.000 ***	1.07	(1) 0.000 ***	1.10	(1) 0.000 ***	1.31	1.26
Centralization14	5.50	(2) 0.906	5.76	(2) 0.831	3.05	(1) 0.457	3.68	(1) 0.325	3.28	3.38
Centralization12	5.57	(2) 0.982	5.85	(2) 0.976	3.33	(1) 0.183	3.97	(1) 0.296	3.59	3.65
Centralization8	5.38	(2) 0.734	5.53	(2) 0.541	3.44	(1) 0.117	4.06	(1) 0.210	4.08	4.18
Centralization6	5.48	(2) 0.882	5.47	(2) 0.434	4.33	(1) 0.000 ***	4.89	(1) 0.001 ***	3.57	3.65

Notes: beta: Statistics based on a smoothed response (see footnote 29 for details). m.a.m: Statistics using mass-at-midpoint assumption. Uncertainty is the standard deviation of a respondent's density forecast and disagreement is the standard deviation of the (mass-at-midpoint) means of the density forecasts. The t-tests assume unequal variance and are one-sided (1) when specified in the hypotheses, two-sided (2) otherwise. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

forecast uncertainty, and the disagreement of respondents. These statistics are calculated using a smoothed response ("beta") and a mass-at-midpoint measure ("m.a.m.").[29]

In the *Shift* treatments, a clear movement of the mean forecasts is observed, in line with Hypothesis 1. Shifting the scale to the right shifts the responses to the right. Shifting the scale to the left shifts the responses to the left. The effect is substantial: The shift amounts to −5.88/−6.40 (beta/mass-at-midpoint) percentage points for *ShiftMinus12* and −1.25/−1.23 for *ShiftMinus4*. In the other direction, we find 1.03/0.96 for *ShiftPlus4* and 2.59/2.47 for *ShiftPlus12*.

In the *Compression* treatments, the entire scale is compressed or expanded. We use forecast uncertainty (the standard deviation of a respondent's density forecast) to have a first glance at Hypothesis 4, which states that the responses compress or expand when we compress or expand the response scale. Compared to *Baseline*, where the uncertainty is 3.08/3.81 (beta, mass at midpoint), uncertainty increases in wide treatments and decreases in narrow treatments. Uncertainty in *Compression4* is 9.83/11.09 and 6.08/7.61 in *Compression2*. In the other direction we find uncertainty of 1.84/2.16 in *Compression0.5* and 1.07/1.10 in *Compression0.25*. Since compressing and expanding the scale also leads to a shift of the responses, we find an indirect knock-on effect on the average mean forecast and on disagreement. When we compress the scale, the average mean forecast is closer to the center of the scale (which is zero in all *Compression* treatments and in *Baseline*), while disagreement is reduced. When we expand the scale, we observe the opposite effect.

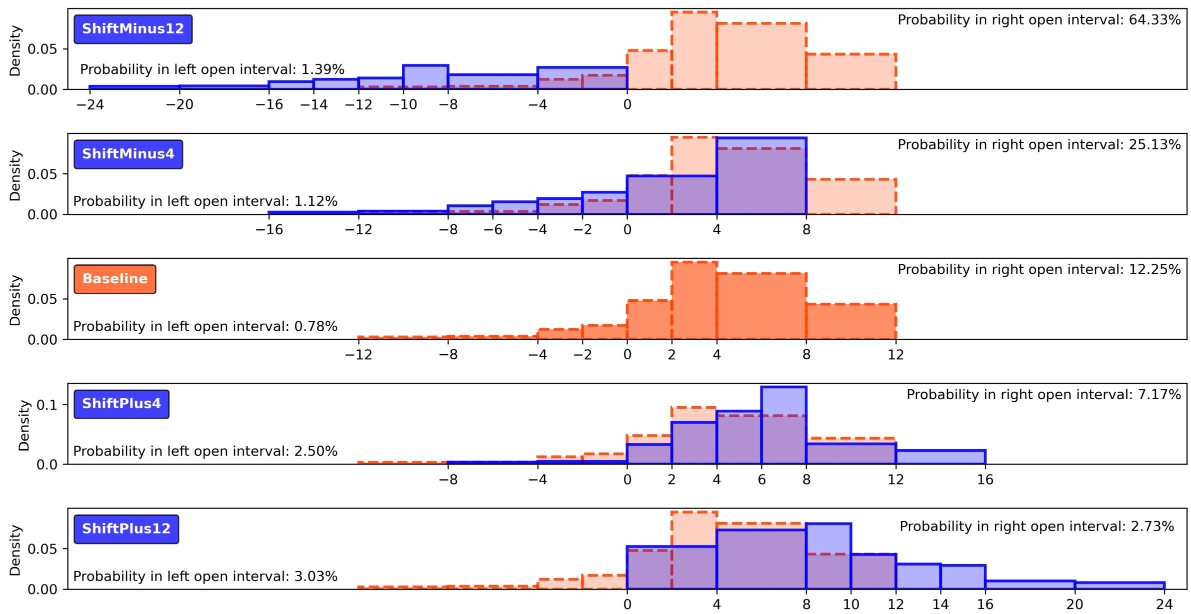
Finally, we find some support for Hypothesis 5 when looking at the uncertainty in treatments with a smaller number of intervals at the center of the scale in the *Centralization* treatments. The uncertainty is 3.44/4.06 in *Centralization8* and 4.33/4.89 in *Centralization6*, however, only the later is significantly different from *Baseline*. In *Centralization12* it is 3.33/3.97 and in *Centralization14* 3.05/3.68 by comparison.

Shift treatments

Figure 3.5 depicts average densities assigned to each interval in the *Shift* treatments, relative to *Baseline*. As the histograms show, the probability mass over the entire scale shifts with the response scale.

[29] Engelberg et al. (2009) suggest to smooth the responses (the histograms) by fitting a parametric distribution from which statistics such as mean, uncertainty, or tail risk may be computed. The procedure assumes a generalized beta distribution when the respondent assigns positive probabilities to three or more intervals and a triangular distribution when the respondent uses one or two intervals. We denote statistics based on this procedure with the abbreviation "beta". This and the mass-at-midpoint procedure require us to make an assumption about the "width" of the open intervals. We assume that the open intervals have twice the width of the adjacent closed interval and when a respondent uses one or both of the open intervals, we follow Engelberg et al. (2009) and treat the limits of the beta distribution as parameters to be estimated. Since most treatments only have small amounts of probability mass in the two open intervals, changing this assumption only leads to small changes of the results. The two important exceptions are *ShiftMinus12* and *Compression0.25* (depicted in gray in Figure 3.4), and care has to be taken when interpreting the figure for these two treatments. Like Armantier et al. (2017), we allow the smooth responses to be bi-modal when respondents supply three or more intervals. See also the discussion about bi-modal responses in Section 3.5. We follow Becker et al. (2022) who extend the original procedure of Engelberg et al. (2009) to response scales with irregular spacing of the intervals.

Figure 3.5: Histograms of the average densities assigned to the intervals in the **Shift treatments** (with Baseline in dashed bars for reference).



Notes: Only closed intervals are illustrated in order to avoid specifying the widths of the open intervals.

This effect can be more clearly illustrated by focusing on one side of the scale. The average probability mass that respondents put into deflation, for example, decreases from 35.67 percent in the *ShiftMinus12* treatment to just 3.11 percent in the *ShiftPlus12* treatment. We test these differences in Table 3.3, which reports the probability masses in the deflation range in comparison to *Baseline*. One-sided Mann–Whitney–Wilcoxon (MWW) tests and t-tests confirm Hypothesis 1 for *ShiftMinus12*, *ShiftMinus4* and *ShiftPlus12*. For *ShiftPlus4* the relocation of probability mass goes in the hypothesized direction but is not significant at the 5% level.[30]

Result 1: *Shifting the response scale leads to a shift of the responses in the same direction.*

We do not find evidence that supports Hypothesis 2 and Hypothesis 3. The treatment interventions in the density forecast do not spill over to the binary inflation/deflation question. When testing Hypothesis 2 for the *Shifting* treatments against *Baseline*, no treatment difference is significant at the 5 percent level.[31]

Result 2: *Shifting the response scale does not affect the responses of the succeeding binary inflation/deflation question.*

Similarly, when testing the point forecasts in the *Shifting* treatments versus *Baseline*, no

[30]See Table A3.3 in Appendix 3.E for additional regressions supporting this pattern and Section 3.4 for significant results regarding treatment *ShiftPlus4* in the larger German survey.

[31]One-sided Fisher exact tests. One treatment difference is significant at the 10% level: *ShiftMinus4* versus *Baseline* (p = 0.056, obs.= 200). It should be noted that very few respondents expected deflation in the binary inflation/deflation question when we conducted the survey in December 2021. In *Baseline*, only a single respondent predicted prices to decline in the following 12 months. All other 100 participants expected prices to increase. In the other *Shift* treatments, the numbers are 95, 93, 94, and 94 (see Table A3.1 in the Appendix).

treatment difference is significant at the 5% level.[32]

Result 3: *Shifting the response scale does not affect the responses of the succeeding question asking for point forecasts.*

Table 3.3: Average probability masses assigned to negative inflation rates (deflation) in the **Shift treatments** (the numbers in the table include the masses assigned to the open intervals).

Treatment	Test Range	Probability Mass			Tests (p-values)			
		Baseline	Treatment	Ratio	MWW		t-Test	
ShiftMinus12	< 0	9.31	35.67	3.83	0.000	***	0.000	***
ShiftMinus4	< 0	9.31	18.35	1.97	0.014	**	0.002	***
ShiftPlus4	< 0	9.31	5.87	0.63	0.230		0.051	*
ShiftPlus12	< 0	9.31	3.03	0.33	0.035	**	0.001	***

Notes: Tests for significant treatment difference (one-sided): MWW (Mann-Whitney-Wilcoxon two-sample statistic) tests, and t-tests (assuming unequal variances). */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Compression treatments

Figure 3.6 shows the average densities assigned to each interval in the *Compression* treatments relative to *Baseline*. Compressing or expanding the scale has a strong effect on the responses and affects mean forecast, forecast uncertainty, and disagreement (see Table 3.2). We now test Hypothesis 4 via changes in the probability mass respondents assign to given ranges of inflation. Since compressing the scale moves interval boundaries, the treatment comparisons require different test ranges. As a rule, we use the largest overlapping range consisting of closed intervals.

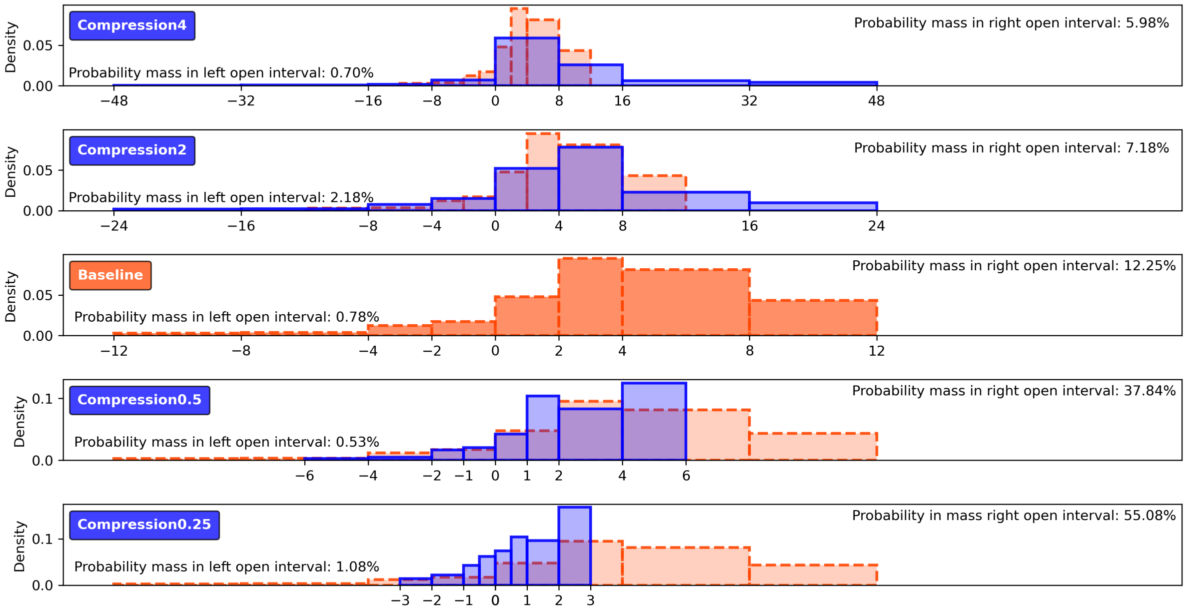
Table 3.4 shows that compressing or expanding the scale significantly compresses and expands the stated responses in treatments *Compression4*, *Compression2*, and *Compression0.25*. When the scale is compressed, respondents move probability mass into intervals covering inflation rates close to zero. When the scale is expanded, respondents move probability mass away from intervals covering inflation rates close to zero.

Result 4: *Compressing or expanding the response scale leads to compressed and expanded responses.*

Result 4 can be explained by a non-responsive use of intervals by the respondents. A responsive participant who tries to accurately "copy" her subjective distribution of inflation expectations onto the response scale would use a different number of intervals in the different *Compression* treatments. As an example, consider a respondent who expects inflation to fall into the range from 0% to 8%. In *Compression4*, this respondent needs only a single interval to express her subjective beliefs. In *Compression2*, the respondent requires two intervals, and in

[32]T-tests: One treatment difference is significant at the 10% level: *ShiftMinus4* vs *Baseline* ($p = 0.085$, obs.= 200). When testing via Mann-Whitney-Wilcoxon (MWW) tests, no treatment difference is significant at the 10% level.

Figure 3.6: Histograms of average densities assigned to the intervals in the **Compression treatments** (with Baseline in dashed bars for reference).



Notes: Only closed intervals are illustrated in order to avoid specifying the widths of the open intervals.

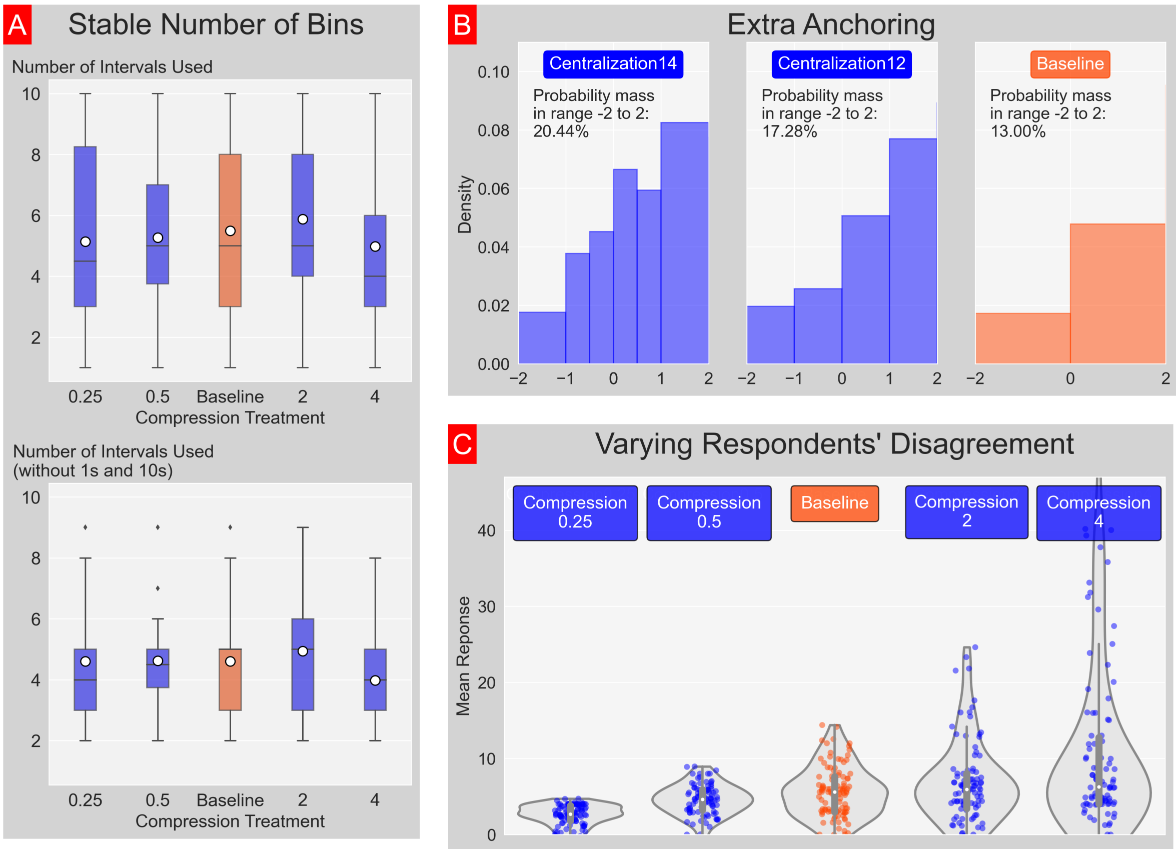
Table 3.4: Average probability masses assigned to overlapping ranges in the **Compression treatments**.

Treatment	**Test Range**	**Probability Mass**			**Tests (p-values)**			
		Baseline	Treatment	Ratio	MWW		t-Test	
Compression4	-8 to 8	68.46	52.71	0.77	0.000	***	0.000	***
Compression2	-8 to 8	68.46	61.23	0.89	0.029	**	0.045	**
Compression0.5	-4 to 4	34.48	36.07	1.05	0.287		0.355	
Compression0.25	-2 to 2	13.00	26.04	2.00	0.000	***	0.000	***

Notes: Tests for significant treatment difference (one-sided): MWW (Mann-Whitney-Wilcoxon two-sample statistic) tests, and t-tests (assuming unequal variances). */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Baseline, 3 intervals are needed. Assuming a responsive use of intervals, one would expect the number of used intervals to decline as the scale gets expanded.

Figure 3.7: Panel A. Boxplots of the number of intervals used by the respondent in the *Compression* treatments, by treatment. Large bright circles indicate averages. Top: All data. Bottom: excluding respondents that use all ten intervals or only a single interval. **Panel B.** Average densities assigned to intervals in the range from −2 to 2 in *Centralization* treatments, by treatment, common axes. **Panel C.** Violin plots and scatterplots (jittered data) of respondents' mean forecasts (mass-at-midpoint) in the *Compression* treatments.



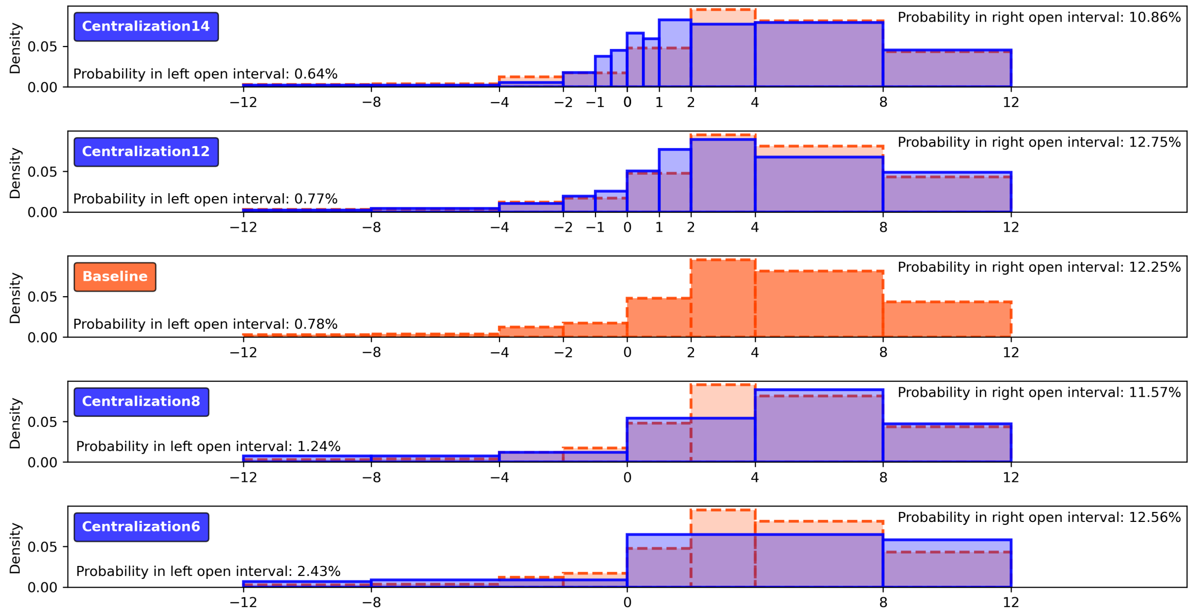
This is not what we find, however. The boxplots in Panel A of Figure 3.7 show that the average number of intervals respondents use is around 6 and does not vary much between treatments. The upper part of the panel includes all data and the bottom part excludes respondents who either use a single interval or use all ten intervals for their answer. The pattern is the same: Respondents tend to use roughly the same number of intervals, independent of the width of the scale. This non-responsive use of intervals may explain the strong treatment effect on respondents' uncertainty (Table 3.2) and may also explain why disagreement declines when we compress the scale (Panel C of Figure 3.7 and Table 3.2).[33]

[33] Studying the Survey of Professional Forecasters (SPF), Glas & Hartmann (2022) find a related effect and report that forecasters do not automatically use twice as many intervals when the interval widths are cut in half. Instead, the forecasters only slightly increase the number of intervals inducing a noticeable drop in uncertainty.

Centralization treatments

Figure 3.8 shows the average densities assigned to each interval in the *Centralization* treatments. As before, we test Hypothesis 5 by comparing probability masses assigned to specific ranges of inflation. The rule we use to select these ranges is to take the smallest central range for which interval boundaries in *Baseline* coincide with the respective treatment boundaries. For *Centralization12* and *Centralization14* the range is from −2 to 2. For *Centralization8*, the range is from −4 to 4 and for *Centralization6* the range is from −8 to 8.

Figure 3.8: Histograms of average densities assigned to the intervals in the **Centralization treatments** (with Baseline in dashed bars for reference).



Notes: Only closed intervals are illustrated in order to avoid specifying the widths of the open intervals.

Table 3.5 shows that it is always the treatment with a higher number of intervals in the comparison range that attracts a higher probability mass. T-tests and MWW tests indicate that for *Centralization14*, *Centralization8*, and *Centralization6*, these treatment differences are significant at least at the 5% level. For *Centralization12*, they are significant at the 10% level.[34]

Result 5: *The probability mass assigned to a given range of inflation rates increases when the response scale provides more intervals in this range.*

The behavior we observe in the *Centralization* treatments has been described in the literature as unpacking bias or partition dependence, and is discussed in detail in Section 3.5. Result 5 highlights how the irregular layout of our response scale in *Baseline* (and in the SCE) allows the unpacking bias to reinforce the central tendency bias. The irregular layout moves probability mass towards the center of the scale, giving the impression that inflation expectations are anchored at low inflation rates. Panel B of Figure 3.7 illustrates this spurious "anchoring" for

[34]See Tables A3.4 to A3.5 in Appendix 3.E for additional regression results supporting this pattern.

Centralization14 and *Centralization12*.

Table 3.5: Average probability masses assigned to overlapping ranges in the **Centralization treatments.**

Treatment	**Test Range**	**Probability Mass**			**Tests (p-values)**			
		Baseline	Treatment	Ratio	MWW		t-Test	
Centralization14	-2 to 2	13.00	20.45	1.57	0.029	**	0.006	***
Centralization12	-2 to 2	13.00	17.28	1.33	0.077	*	0.055	*
Centralization8	-4 to 4	34.48	26.56	0.77	0.042	**	0.028	**
Centralization6	-8 to 8	68.46	58.90	0.86	0.003	***	0.011	**

Notes: Tests for significant treatment difference (one-sided): MWW (Mann-Whitney-Wilcoxon two-sample statistic) tests, and t-tests (assuming unequal variances). */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Respondents' internal consistency

In order to test Hypothesis 6, we follow Engelberg et al. (2009) and construct nonparametric bounds on the mean and median of the histograms. We then examine whether the reported point forecasts fall into the bounds. The procedure does not impose specific distributional assumptions on the underlying densities.[35] Table A3.1 in Appendix 3.C shows average point forecasts for all treatments.

For each respondent, we place the probability mass the respondent assigns to an interval at the interval's lower or upper limits. Doing this for each interval of the response scale and summing up, we obtain lower and upper bounds on a respondent's mean. If the point forecast falls within those bounds, it is consistent with the mean. To construct the lower and upper bounds on the median, let $j \in \{1, 2, ..., N\}$ denote the index of the response intervals whose lower bounds we denote θ_j and whose upper bounds we denote θ_{j+1}. With p_{ij}, the probability assigned to interval j by respondent i, the point forecast must fall within the interval $[\theta_k, \theta_{k+1}]$, where k is determined by $\sum_{s=1}^{k} p_{is} \leq 0.5$ and $\sum_{s=1}^{k+1} p_{is} \geq 0.5$, to be consistent with the median.

As a reference, we also calculate the consistency measures for the SCE for the December 2021 wave. Table 3.6 shows the results of the consistency tests for all 13 treatments and for the SCE. Respondents in *Baseline* display the same consistency as the respondents of the SCE. When we compare the *Shift* treatments with *Baseline*, only *ShiftMinus12* shows a significant difference for the mean, though not for the median.

Table 3.6 also reports the results for the *Compression* and *Centralization* treatments. Some caution should be used, however, when interpreting these results. Compressing, expanding, shifting, splitting, or combining intervals changes the "consistency target" (since not all intervals are equally wide). Wider targets (bounds) are easier to hit. It is therefore not surprising if the share of respondents with consistent answers grows in *Compression4* and *Centralization6*. In

[35] Several papers study respondents' internal consistency by comparing point forecasts with measures of central tendency derived from the subjective probability distribution. For household surveys see Zhao (2022), Delavande & Rohwedder (2011), Bruine de Bruin et al. (2011), and for surveys of professionals see Engelberg et al. (2009) and Clements et al. (2023) among others.

Table 3.6: Consistency. Shares of point forecasts that fall within the bounds on the mean (column 2) or the median (column 4) of the density forecasts, by Treatment and using data from the SCE.

Treatment	**Statistics**				
		Mean-consistent		Median-consistent	
	Observations	Share	P-value	Share	P-value
Baseline	101	0.62		0.63	
SCE (December 2021)	1283	0.58	0.403	0.62	0.915
ShiftMinus12	99	0.45	0.023**	0.52	0.115
ShiftMinus4	99	0.60	0.772	0.64	1.000
ShiftPlus4	98	0.55	0.316	0.53	0.153
ShiftPlus12	98	0.55	0.316	0.63	1.000
Compression4	99	0.73	0.133	0.73	0.174
Compression2	99	0.62	1.000	0.66	0.769
Compression0.5	96	0.58	0.662	0.61	0.883
Compression0.25	100	0.39	0.001***	0.51	0.088*
Centralization14	96	0.66	0.659	0.64	1.000
Centralization12	96	0.58	0.662	0.54	0.196
Centralization8	99	0.68	0.461	0.71	0.295
Centralization6	99	0.78	0.021**	0.75	0.094*

Notes: All data from December 2021. Two-sided Fisher Exact tests compared to Baseline. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Compression4, for example, a single interval covers the entire range from 0% to 8%. Section 3.5 continues this discussion.

Result 6: *Between* 39.0% *and* 77.8% *of respondents report consistent answers.*

Impact of respondents' proficiency

After each question about inflation expectations, we ask subjects to state their subjective certainty for this answer. Respondents then complete a questionnaire with questions on financial literacy, highest obtained degree, and knowledge of the Fed's inflation target. We refer to the collection of these measures as "proficiency". To evaluate whether respondents with higher proficiency are less affected by changes of the responses scale (Hypothesis 7), we use the inflation ranges established in the preceding subsections and regress the probability mass assigned to the ranges on the proficiency variables, treatment dummies, interaction terms and several control variables. Specification (3) of Tables A3.3 to A3.6 in Appendix 3.E reports the results.

The financial literacy interaction is never significant at the 5% level for any *Shift* or *Centralization* treatment. It is significant at the 1% level for *Compression4* and at the 5% level for *Compression0.5*. The interaction term for knowledge of the inflation target is never significant at the 1% level for any treatment and significant at the 5% level only for *Compression0.25*. Having high education leads to significant interactions at the 5% level only for the treatment *ShiftMinus12*. Overall, we find little evidence for Hypothesis 7.

Result 7: *There is little evidence that higher educated or more knowledgeable respondents are* *affected less by changes of the response scale.*

According to Hypothesis 8, respondents with higher proficiency should be more certain in their answers. To test this, we regress respondents' certainty on the proficiency variables and other controls in Table A3.7 in Appendix 3.E. Knowing the inflation target makes respondents more certain of their answer in all three forecasts (density forecast, point prediction, binary inflation/deflation forecast). However, for the point prediction, this becomes insignificant when controls are added. Instead, respondents become less certain here with higher financial literacy. Education never has a significant influence on subjective certainty. For all three questions, the higher a respondent's forecast, the higher their certainty. Women are always less certain, Republicans are always more certain.

Result 8: *Respondents who know the Federal Reserve Bank's inflation target are more certain in their forecasts. However, higher reported education or financial literacy do not increase respondents' certainty.*

3.4 The German survey

In addition to the data collected for the US via Prolific, we included two treatments, *ShiftPlus4* and *Centralization14*, in the Bundesbank Online Panel Households (BOP-HH) in June 2022. The BOP-HH closely follows the SCE in its design of the inflation density question, only the order of the intervals differs. The response scale of the BOP-HH starts with deflation whereas the response scale of the SCE starts with inflation (see question 1 in Appendix 3.A).[36] Year-on-year CPI inflation in Germany was reported to be 7.1 percent in June 2022 and of similar magnitude in the preceding months (Bundesbank, 2022).

3.4.1 Implementation

In June 2022, 4460 German households participated in Wave 30 of the BOP-HH.[37] We removed observations from the sample whenever a household did not report probabilistic inflation expectations or if information for any of the socioeconomic characteristics is missing. We also exclude the response from one household which did not answer the question of whether she expects inflation or deflation. This leaves $4,094$ observations in our sample for Wave 30. Of these, 1356 participated in the standard BOP-HH (*Baseline*) question, 1377 in *ShiftPlus4*, and 1361 in *Centralization14*.

3.4.2 Results

In Table 3.7, we replicate the analysis of Table 3.2 for the German data. The predicted treatment differences go in the same direction as in the US data. The differences are highly significant for

[36] For a technical description of the BOP-HH Survey see Beckmann & Schmidt (2020).

[37] In Becker et al. (2023) we use the panel structure of the survey and compare the June wave with the preceding and the subsequent waves.

Table 3.7: Treatment differences for the German survey.

Treatment	Statistics										
	Mean Forecast				Forecast Uncertainty				Disagree-ment	
	beta		m.a.m.		beta		m.a.m.		beta	m.a.m.
Name	Avg.	P-value	Avg.	P-value	Avg.	P-value	Avg.	P-value		
Baseline	6.63		6.72		2.12		2.18		4.01	4.06
ShiftPlus4	7.22	(1) 0.000 ***	7.28	(1) 0.000 ***	1.79	(2) 0.000 ***	1.89	(2) 0.000 ***	3.55	3.59
Centralization14	6.42	(2) 0.146	6.50	(2) 0.162	2.04	(1) 0.079 *	2.07	(1) 0.066 *	3.88	3.86

Notes: beta: Statistics based on a smoothed response. See footnote 29 for details. m.a.m.: Statistics using mass-at-midpoint assumption. Uncertainty is the standard deviation of a respondent's forecast and disagreement is the standard deviation of respondents' mean forecasts. T-tests assume unequal variance and are one-sided, (1), when specified in the hypotheses, two-sided, (2), otherwise. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Table 3.8: Shift and centralization treatments in the German survey.

Treatment	Test Range	Probability Mass			Tests (p-values)	
		Baseline	Treatment	Ratio	MWW	t-Test
ShiftPlus4	< 0	7.10	3.72	0.52	0.000 ***	0.000 ***
Centralization14	-2 to 2	5.68	8.78	1.54	0.007 ***	0.000 ***

Notes: Tests for significant difference (one-sided) in average probability masses. MWW (Mann-Whitney-Wilcoxon two-sample statistic) and t-test (assumes unequal variances). */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

t-tests of the *ShiftPlus4* treatment differences and weakly significant for the *Centralization14* treatment. As in Section 3.3, we also employ tests that directly use the probability masses assigned to the intervals. Table 3.8 repeats the analysis of Tables 3.3 and 3.5 for the German data. We find significant differences for both treatments. The size of the treatment effects is surprisingly similar to the US data results. The ratio of the probability mass in the deflation region of *ShiftPlus4* is 0.52 times that of the probability mass in the deflation region of *Baseline* in Germany. In the US data, this ratio is 0.63. For *Centralization14*, the probability mass in the range from -2 to 2 is 1.54 times that of *Baseline* in Germany. In the US, this ratio is 1.57. Overall, despite running the treatments in a different country and at different times, we find the same direction of treatment effects and very similar effect sizes.

3.5 Discussion

One way to interpret the results of the two previous sections is to assume non-rational behavior via behavioral biases. An alternative interpretation that does not presuppose non-rationality can be found in Bayesian updating. In the first part of this section we describe the two interpretations in more detail. The second part outlines steps that could mitigate the problem, and discusses what to do in times of high inflation when respondents assign large probability masses to the open intervals. To keep the presentation simple, we refer to any discrepancy between a respondent's true or prior beliefs and the measured beliefs as “measurement bias”.

3.5.1 Interpretation of the measurement bias

One way to interpret our results is to assume non-rational behavior. Instead of maintaining coherent probability distributions over future events (such as inflation) and following rational updating rules such as Bayes’ law, respondents might follow simpler heuristics. Following this line of reasoning, the treatment differences we find in Results 1, 4 and 5 can be explained by behavioral biases that are known in the literature from other settings.

The central tendency bias (Hollingworth, 1910; Duffy et al., 2010) refers to respondents’ propensity to prefer answers in the middle of the response scale. This could explain, as seen in Result 1, why respondents shift their reported probability distributions following a shift of the response scale. In Result 4 we observe that respondents tend to assign probability masses to the intervals without properly taking into account the compression of the scale. This is in line with support theory (Tversky & Koehler, 1994) and with partition dependence (Fox & Rottenstreich, 2003). Finally, we find that respondents tend to assign a larger amount of probability mass to a given range of inflation rates, the more intervals the scale uses to represent this range (Result 5). This is similar to behavior found in other studies where it is referred to as unpacking bias (Tversky & Koehler, 1994; Sonnemann et al., 2013). One piece of evidence favoring an explanation via behavioral biases is the lack of knock-on effects of the treatment intervention onto the binary inflation/deflation question and the point forecast (Results 2 and 3).

A second interpretation of our results is that the treatment differences described above are the result of a rational cognitive process in which respondents use two sources of information when providing an answer. The first source of information is the respondent’s prior knowledge about future inflation, based on information about past or current inflation, possibly combined with information about the macro-economic environment and the central bank’s policy. The second source of information is what is called *context* in the survey research literature (see Schuman, 1992; Schwarz, 2010). Context includes any information respondents obtain from participating in the survey. In the case of density questions, the response scale is an important part of the question’s context. When asked about their inflation expectations, respondents may consider the response scale to reflect the surveyor’s (i.e. the central bank’s) own expectations. For example, by putting certain values of inflation in the center of the scale, the central bank signals that these values are more plausible than values in the peripheral intervals.[38] Evidence favoring the rational updating interpretation comes from an asymmetry of the treatment differences in

[38] In the US survey, we self-identified as researchers from the University of Heidelberg. The German survey was conducted by the Bundesbank. Apart from the response scale, respondents may also extrapolate information from the wording of a question (Schuman & Presser, 1996), the order of a question (Phillot & Rosenblatt-Wisch, 2018), or the affiliation of the surveying researcher (Schwarz, 2010).

Result 1.[39] While a behavioral bias, if linear, would work similarly in both the *ShiftPlus* and *ShiftMinus* treatments, updating can explain an asymmetry via priors that are not centered on zero.

3.5.2 Mitigating the measurement bias

It is probably natural that taking time-differences comes to mind when looking for a way to mitigate the measurement bias. A measurement bias that is constant over time would cancel out when differencing. However, the assumption of a measurement bias that is constant over time seems improbable, as the example below illustrates.

Let $\mu_{i,t}$ be respondent i's mean forecast in period t and let $\mu^*_{i,t}$ be respondent i's prior (or true) inflation expectations in t. Assuming that the measurement bias ($\beta_{i,t}$) enters additively, we can express the change in inflation expectations as

$$(\mu_{i,t} - \mu_{i,t-1}) = \left(\mu^*_{i,t} - \mu^*_{i,t-1}\right) + (\beta_{i,t} - \beta_{i,t-1})\,.$$

To some extent, taking differences could alleviate the measurement bias, so that $(\beta_{i,t} - \beta_{i,t-1})$ is approximately zero. It seems possible, for example, that gender effects on $\beta_{i,t}$ are time invariant. But in general, $\beta_{i,t}$ is likely to vary with time because $\beta_{i,t}$ itself depends on the prior expectations $\mu^*_{i,t}$. Figure 3.9 illustrates this dependence. Result 1 shows that respondents are drawn towards the center of the scale, so the observed responses will typically lie between the prior expectations and the center of the scale. However, differencing mitigates the measurement bias only if the bias remains constant when the prior expectations vary over time. But when $\mu_{i,t}$ is bounded by the scale center, $\beta_{i,t}$ will decrease as $\mu^*_{i,t}$ approaches the scale center (Panel A). When the prior expectations happen to fall on the other side of the center (Panel B), the measurement bias will even change sign.[40] The assumption that the measurement bias is time-invariant is, therefore, not very convincing. Taking differences does not yield reliable estimates of a respondent's true changes in expectations.

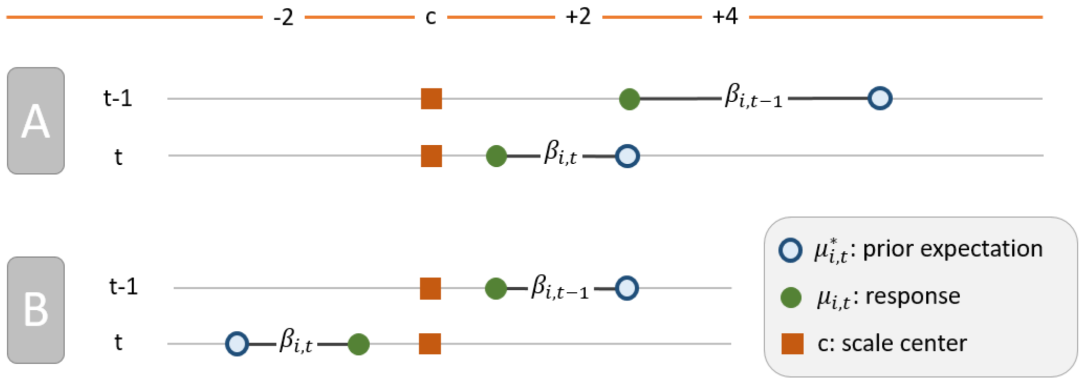
As the measurement bias is in part introduced by the survey itself, a more promising approach to mitigate it is to modify the design of the question. One possible change is to use regularly-spaced response scales. Making the intervals narrow in some range is often motivated by the desire to give respondents the possibility to be more specific in some range while keeping the overall number of intervals reasonably small.[41] However, as the results in Sections 3.3 and 3.4

[39] We use the fitted (beta) means of the *Shift* treatments to calculate the difference between individual means and the average of means in *Baseline*. Then we test whether the difference of *Baseline* and *ShiftPlus4* is different from the difference of *ShiftMinus4* and *Baseline* via t-tests (and similar for the *ShiftPlus12* and *ShiftMinus12* treatments). The differences for the +/-12 treatments are significantly different ($p \leq 0.001$, obs. = 197), but the differences for the +/-4 treatments are not ($p = 0.694$, obs. = 197).

[40] The sign change of $\beta_{i,t}$ may open the possibility to identify bounds on the true expectations. We illustrate this idea in Appendix 3.G.

[41] Here we are referring to an irregular spacing of the response scale in a range that respondents consider probable. There is a related problem with the two open intervals but the survey questions are typically designed in a way that keeps the probability mass in the open intervals small.

Figure 3.9: Illustration of an additive, time-varying measurement bias. In Panel B, $\beta_{i,t}$ reverses its sign from time $t-1$ to time t.



show, the narrow intervals attract additional probability masses, giving the spurious impression that values in the narrow intervals are expected more often.

The irregular spacing has other consequences. The first is that the consistency bounds are tighter when the intervals are narrow, making it more difficult for respondents to provide consistent responses (Zhao, 2022). In a survey with an irregularly spaced scale, such as *Baseline*, respondents expecting high inflation will then appear more consistent than respondents expecting low inflation.[42] A second consequence concerns the shape of the response. A response with a single mode (peak) is often a desirable property and, in fact, uni-modality is the "most basic assumption" in the parametric analysis of Engelberg et al. (2009, p. 36). Because of the irregular spacing, the subjective densities may be bi-modal even though the underlying probabilities are single peaked.[43]

A second promising design change is to give each respondent a personalized response scale centered on the respondent's point forecast.[44] This design minimizes the impact of the central tendency bias and reduces the need to provide very wide scales, rendering the irregular scales (introduced to achieve precise results near the assumed center of the distribution, while still allowing a broad range of expectations) unnecessary.

Such a design also eliminates the need to adjust the response scale in times of high inflation.

[42] For example, the bounds on the mean for a respondent in *Baseline* who expects inflation to fall into the range from 4 to 8 percent are twice as wide as the bounds for a respondent who expects inflation to fall into the range from 0 to 4.

[43] In the US survey, a large majority of the respondents supplies uni-modal responses (see Table A3.1 in Appendix 3.C). But there are 112 (out of 1279) responses whose densities are bi-modal even though the bar-chart of the probabilities is uni-modal (the opposite occurs 22 times). As an example, consider respondent with id 659 who assigns single-peaked probabilities of 10, 15, 45, and 30 percent to the intervals 7 to 10 of the *Baseline* treatment. Since interval 7 is only half as wide as interval 8, the subjective histogram has two modes, and it is unclear whether this bi-modality is intentional.

[44] Dominitz & Manski (1997) use a scale determined by preliminary questions about subjective lowest and highest outcomes while studying household incomes, yet they warn against using these answers as minima and maxima of the scale. The Survey of Expectations of the Central Bank of the Republic of Turkey (CBRT) uses a regularly-spaced response scale centered on the respondents' point forecast, see e.g., Gülşen & Kara (2019). Crosetto & De Haan (2022) go a step further by letting respondents essentially construct their own scale via a click-and-drag interface.

These adjustments are necessary when respondents assign large probability masses in the open intervals. For example, the Survey of Professional Forecasters (SPF) of the Federal Reserve Bank of Philadelphia has regularly adjusted its response scale in the past decades by adding or removing intervals. The disadvantage of this approach is that any of these adjustments is likely to affect the responses. Responses from before and after the adjustment are, therefore, not directly comparable. A possible way to alleviate this problem could be to split the survey population and run two surveys (with the new and the old scale) in parallel for some time gathering data that could allow a chaining of the two series.[45]

3.6 Conclusion

In the past decade, several major central banks followed the New York Fed and started to elicit households' inflation expectations via density questions. An often cited advantage of density questions is that they allow us not only to quantify mean and median forecasts but also other variables that are valuable for central banking such as respondents' uncertainty or their perceived tail risk. Using the original question of the New York Fed as our baseline, the experiments in this paper provide a thorough test of how measured beliefs (the reported inflation forecasts) vary when we vary the response scale. The results show that shifting, compressing or expanding the scale leads to shifted, compressed and expanded forecasts. Beliefs measured using a density question systematically depend on the response scale. The resulting measurement bias is substantial, indicating that the quantitative nature of inflation density forecasts is deceptive. As such, inflation density forecasts can provide misleading information about how well respondents' expectations are anchored at a certain value. The measurement bias can vary over time so that even in differences, the forecasts are only suggestive.

However, the experiments also show that the measurement bias can be explained by well-known behavioral biases or even be rationalized. Understanding the underlying causes is a first step to control the measurement bias. Providing each respondent with a personalized response seems a promising way forward. Moreover, our experiments focused on households and it is possible that firms and especially professional forecasters are less affected by the behavioral biases than households, but it would be good to see more research in this direction.

[45] The SCE follows a different approach, using a comparatively wide response scale, and no change to the scale was considered necessary so far. Still, in March 2022 respondents assigned more than a fifth of the probability mass in the upper open interval (the average probability mass in the upper open interval between 2017 and 2019 is less than seven percent).

Chapter 3 References

Armantier, O., Topa, G., Van der Klaauw, W., & Zafar, B. (2017). An overview of the survey of consumer expectations. *Economic Policy Review*(23-2), 51–72.

Becker, C., Duersch, P., Eife, T. A., & Glas, A. (2022). Extending the procedure of engelberg et al.(2009) to surveys with varying interval-widths. *AWI Discussion Paper No. 707*.

Becker, C., Duersch, P., Eife, T. A., & Glas, A. (2023). Households' probabilistic inflation expectations in high-inflation regimes. *FAU Discussion Papers in Economics 1_2023*.

Beckmann, E., & Schmidt, T. (2020). *Bundesbank online pilot survey on consumer expectations* (No. January 2020). Frankfurt (Main). (Technical Paper)

Bruine de Bruin, W., Chin, A., Dominitz, J., & van der Klaauw, W. (2023). Household surveys and probabilistic questions. In *Handbook of economic expectations* (pp. 3–31). Elsevier.

Bruine de Bruin, W., Manski, C. F., Topa, G., & Van Der Klaauw, W. (2011). Measuring consumer uncertainty about future inflation. *Journal of Applied Econometrics*, *26*(3), 454–478.

Bruine de Bruin, W., Van der Klaauw, W., Topa, G., Downs, J. S., Fischhoff, B., & Armantier, O. (2012). The effect of question wording on consumers' reported inflation expectations. *Journal of Economic Psychology*, *33*(4), 749–757.

Bundesbank. (2022). Monthly report. *Deutsche Bundesbank*(06/2022).

Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Clements, M. P., Rich, R. W., & Tracy, J. S. (2023). Surveys of professionals. In *Handbook of economic expectations* (pp. 71–106). Elsevier.

Coibion, O., Gorodnichenko, Y., & Kumar, S. (2018). How do firms form their expectations? new survey evidence. *American Economic Review*, *108*(9), 2671–2713.

Coibion, O., Gorodnichenko, Y., Kumar, S., & Pedemonte, M. (2020). Inflation expectations as a policy tool? *Journal of International Economics*, *124*, 103297.

Coibion, O., Gorodnichenko, Y., & Weber, M. (2022). Monetary policy communications and their effects on household inflation expectations. *Journal of Political Economy*, *130*(6).

Crosetto, P., & De Haan, T. (2022). Comparing input interfaces to elicit belief distributions. *GAEL Working Paper*(01/2022).

Delavande, A., & Rohwedder, S. (2011). Individuals' uncertainty about future social security benefits and portfolio choice. *Journal of Applied Econometrics*, *26*(3), 498–519.

Dominitz, J., & Manski, C. F. (1997). Using expectations data to study subjective income expectations. *Journal of the American Statistical Association*, *92*(439), 855–867.

Duffy, S., Huttenlocher, J., Hedges, L. V., & Elizabeth Crawford, L. (2010). Category effects on stimulus estimation: Shifting and skewed frequency distributions. *Psychonomic Bulletin & Review*, *17*(2), 224–230.

Engelberg, J., Manski, C. F., & Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business & Economic Statistics*, *27*(1), 30–41.

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, *14*(3), 195–200.

Glas, A., & Hartmann, M. (2022). Uncertainty measures from partially rounded probabilistic forecast surveys. *Quantitative Economics*, *13*(3), 979–1022.

Gülşen, E., & Kara, H. (2019). Measuring inflation uncertainty in turkey. *Central Bank Review*, *19*(2), 33–43.

Hollingworth, H. L. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, *7*(17), 461–469.

Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of economic literature*, *52*(1), 5–44.

Manski, C. F. (2018). Survey measurement of probabilistic macroeconomic expectations: Progress and promise. *NBER Macroeconomics Annual*, *32*, 411-471.

Payne, S. L. (Ed.). (1951). *The art of asking questions.* Princeton: Princeton University Press.

Phillot, M., & Rosenblatt-Wisch, R. (2018). *Inflation expectations: The effect of question ordering on forecast inconsistencies* (Tech. Rep.). Swiss National Bank.

Schuman, H. (1992). Context effects: State of the past/state of the art. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 5–20). New York, NY: Springer New York.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context.* Sage.

Schwarz, N. (2010). Measurement as cooperative communication: What research participants learn from questionnaires. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The sage handbook of measurement* (pp. 43–61). Los Angeles: Sage.

Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, *49*(3), 388–395.

Sonnemann, U., Camerer, C. F., Fox, C. R., & Langer, T. (2013). How psychological framing affects economic market prices in the lab and field. *PNAS*, *110*(29), 11779–11784.

Sudman, S., & Bradburn, N. (Eds.). (1974). *Response effects in surveys: A review and synthesis.* Chicago: Aldine.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567.

Zhao, Y. (2022). Internal consistency of household inflation expectations: Point forecasts vs. density forecasts. *International Journal of Forecasting*.

Appendix 3

3.A Instructions of the US survey

Annotations in italics represent comments on formatting/coding.

Introduction

Welcome!

You will take part in an academic survey conducted by the University of Heidelberg, Germany. We are interested in your personal views regarding the future inflation rate: it is therefore important that you answer honestly and read the questions very carefully before answering. This survey should take (on average) less than **8 minutes** to complete. For completing this survey, you will receive a fixed payment of **£1.00 (approximately $*[current value in US dollar]*)**.

Participation in this survey is entirely voluntary and you will remain anonymous throughout the survey. Results may include summary data, but you will never be identified. By continuing, you consent to the publication of survey results. Note that you cannot save and come back later to answer the survey. If you have any questions regarding this survey, you may contact us at **survey2021@awi.uni-heidelberg.de**.

If you understand and agree to the above information, please check "I consent, begin survey" below and click "Next" to begin. Otherwise, check "I do not consent" below and click "Next" to not take part in the survey.

O I consent, begin study

O I do not consent

[Next]

Instructions

We want to learn about your current outlook for future inflation in the United States. To do so, we will ask you a couple of questions. We are interested in your views and opinions. Your responses are confidential, and it helps us a great deal if you respond as carefully as possible. If you should come to any question that you can't or don't want to answer, just click on Next until the next question appears.

In some of the following questions, we will ask you to think about the percent chance of something happening in the future. Your answers can range from 0 to 100, where 0 means there is absolutely no chance, and 100 means that it is absolutely certain.

Thank you for your participation!

[Next]

Question 1

We would like you to think about the different things that may happen to inflation over the next 12 months. We realize that this question may take a little more effort.

In your view, what would you say is the percent chance that, over the next 12 months...

the rate of inflation will be 12% or higher		percent chance
the rate of inflation will be between 8% and 12%		percent chance
the rate of inflation will be between 4% and 8%		percent chance
the rate of inflation will be between 2% and 4%		percent chance
the rate of inflation will be between 0% and 2%		percent chance
the rate of deflation (opposite of inflation) will be between 0% and 2%		percent chance
the rate of deflation (opposite of inflation) will be between 2% and 4%		percent chance
the rate of deflation (opposite of inflation) will be between 4% and 8%		percent chance
the rate of deflation (opposite of inflation) will be between 8% and 12%		percent chance
the rate of deflation (opposite of inflation) will be 12% or higher		percent chance
Total		**percent chance**

[Next]

Notes:

1. *Bin labels shown here are taken from the Baseline condition.*
2. *Page includes a running total that is updated as soon as a participant enters a value into one of the bins.*

Error messages for this page:

1. *Upon submitting an empty forecast (total of 0 percent chance):*
   Your answers are important to us. Please provide an answer even if you are not sure. Otherwise click ***Next*** *to continue.*
2. *Upon submitting a forecast with a total of less than 100 percent:*
   Your total adds up to [percent sum]%. Please change the numbers in the table so they add up to 100%. Otherwise click ***Next*** *to continue.*

Question 2

How certain do you feel about your response to the previous question?

O Very Certain
O Certain
O Somewhat Certain
O Somewhat Uncertain

O Uncertain
O Very Uncertain
[Next]

Question 3

Over the next 12 months, do you think that there will be inflation or deflation? (Note: deflation is the opposite of inflation)

Please choose one.
O Inflation
O Deflation (the opposite of inflation)
[Next]

Error messages for this page:

1. *Upon trying to submit the page without providing an answer:*
   Your answers are important to us. Please provide an answer even if you are not sure. Otherwise click ***Next*** *to continue.*

Question 4

How certain do you feel about your response to the previous question?
O Very Certain
O Certain
O Somewhat Certain
O Somewhat Uncertain
O Uncertain
O Very Uncertain
[Next]

Question 5

What do you expect the rate of *[inflation/deflation]* to be **over the next 12 months**? Please give your best guess.

Over the next 12 months, I expect the rate of *[inflation/deflation]* to be ☐ %

[Next]

Notes:

1. *Whether inflation or deflation is shown in the text depends on the participants' answer to Question 3. If participants answered in Question 3 that they expect inflation over the next 12 months, inflation was shown in Question 5 and deflation otherwise.*

Error messages for this page:

1. *Upon trying to submit the page without providing an answer:*
   *Your answers are important to us. Please provide an answer even if you are not sure. Otherwise click **Next** to continue.*

Question 6

How certain do you feel about your response to the previous question?

- O Very Certain
- O Certain
- O Somewhat Certain
- O Somewhat Uncertain
- O Uncertain
- O Very Uncertain

[Next]

Questionnaire

To conclude the survey, we would like to ask you some questions about you and your household.

Age (leave blank if you prefer not to tell): [ ]

Gender:

- O Prefer not to answer
- O Female
- O Male
- O Other

Highest educational degree:

- O Prefer not to answer
- O High school diploma
- O Some college no degree
- O Associate's degree occupational
- O Associate's degree academic
- O Bachelor's degree
- O Master's degree
- O Professional degree
- O Doctoral degree

Please select "Squirrel". This question just helps us to screen out random clicking:

O Prefer not to answer
O Elephant
O Capybara
O Wolf
O Squirrel
O Mouse

The US Federal Reserve System (Fed) tries to control the inflation rate by keeping it close to a specific target value. What do you think is this target for the inflation rate?

O Prefer not to answer
O Positive inflation that averages 2% over time
O Negative inflation that averages -2% over time
O Positive inflation that averages 1% over time
O On average zero inflation over time
O Don't know

Your political orientation:

O Prefer not to answer
O Republican
O Democrat
O Independent
O Other

State of residence: ☐ *drop-down menu with list of states*

Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?

O Prefer not to answer
O More than $102
O Exactly $102
O Less than $102
O Don't know

Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, with the money in this account, would you be able to buy...

O Prefer not to answer
O More than today
O Exactly the same as today
O Less than today
O Don't know

Do you think the following statement is true or false?
"Buying a single company stock usually provides a safer return than a stock mutual fund."

O Prefer not to answer
O True
O False
O Don't know
[Next]

Error messages for this page:

1. *When not submitting an answer to one of the questions:*
   Your answers are important to us. Please provide an answer or select "Refuse to answer".

End page

Thank you for your participation!

If you have any questions regarding this survey, you may contact us at **survey2021@awi.uni-heidelberg.de**.

Click here to confirm your participation and to return to Prolific. *[Sentence is hyperlink]*

No consent given page

As you do not wish to participate in this study, please return your submission on Prolific by selecting the 'Stop without completing' button.

If you have any questions regarding this study, you may contact us at **survey2021@awi.uni-heidelberg.de**.

You can close this window now.

Timeout page

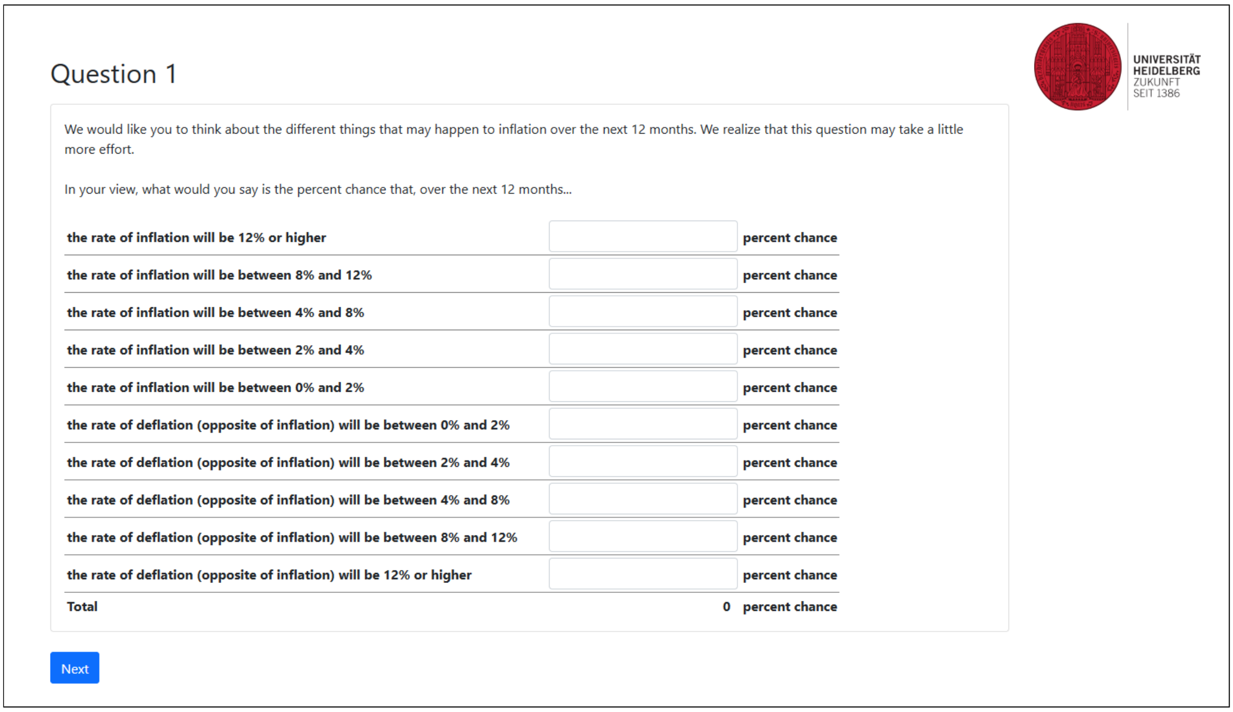
You did not complete the page in time. Thus you cannot finish this assignment.

If you have any questions regarding this study, you may contact us at **survey2021@awi.uni-heidelberg.de**.

You can close this window now.

3.B Screenshots of the US survey

Figure A3.1: Screenshot of the density question for the Baseline condition.


Question 1

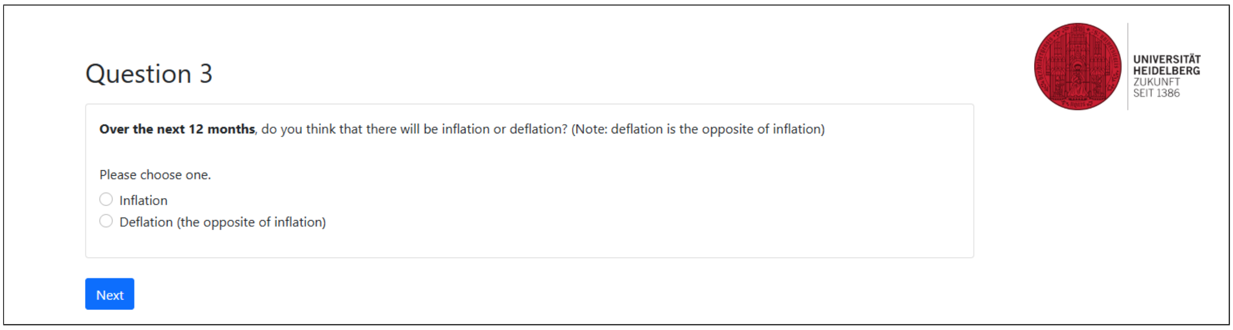
We would like you to think about the different things that may happen to inflation over the next 12 months. We realize that this question may take a little more effort.

In your view, what would you say is the percent chance that, over the next 12 months...

the rate of inflation will be 12% or higher		**percent chance**
the rate of inflation will be between 8% and 12%		**percent chance**
the rate of inflation will be between 4% and 8%		**percent chance**
the rate of inflation will be between 2% and 4%		**percent chance**
the rate of inflation will be between 0% and 2%		**percent chance**
the rate of deflation (opposite of inflation) will be between 0% and 2%		**percent chance**
the rate of deflation (opposite of inflation) will be between 2% and 4%		**percent chance**
the rate of deflation (opposite of inflation) will be between 4% and 8%		**percent chance**
the rate of deflation (opposite of inflation) will be between 8% and 12%		**percent chance**
the rate of deflation (opposite of inflation) will be 12% or higher		**percent chance**
Total	**0**	**percent chance**

Next

Figure A3.2: Screenshot of the inflation/deflation question.


Question 3

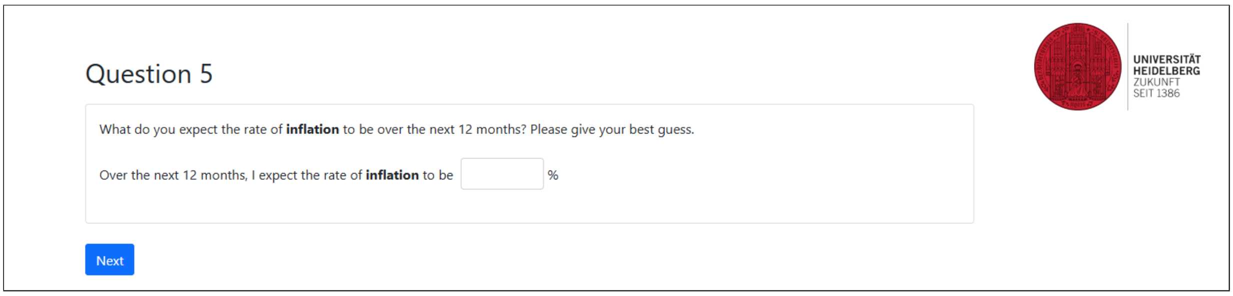
Over the next 12 months, do you think that there will be inflation or deflation? (Note: deflation is the opposite of inflation)

Please choose one.

- Inflation
- Deflation (the opposite of inflation)

Next

Figure A3.3: Screenshot of the point forecast.


Question 5

What do you expect the rate of **inflation** to be over the next 12 months? Please give your best guess.

Over the next 12 months, I expect the rate of **inflation** to be %

Next

3.C Descriptive statistics

Table A3.1: Descriptive statistics by treatment (US survey).

Treatment					**Results**											
		Response scale			Point forecast					Responses with				Uni-modality		
	obs	#	Center	Span	Mean	Trimmed mean	Median	Mean forecast	Intervals used	single interval	full set	open intervals	gaps	Probability	Density	Expecting inflation
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Baseline	101	10	0	24	8.30	6.56	6.00	5.56	6.35	1	28	84	5	87	82	100
ShiftMinus12	99	10	-12	24	7.52	6.01	5.00	-0.32	5.28	19	27	90	3	85	71	95
ShiftMinus4	99	10	-4	24	5.88	6.01	6.00	4.31	6.20	2	37	98	4	73	70	93
ShiftPlus4	98	10	4	24	7.50	6.88	7.00	6.59	7.10	1	32	71	6	74	69	94
ShiftPlus12	98	10	12	24	8.72	7.38	7.00	8.16	7.51	2	25	74	0	90	86	94
Compression4	99	10	0	96	11.83	9.59	7.00	10.98	6.38	1	38	60	1	86	83	99
Compression2	99	10	0	48	9.39	8.04	7.00	6.23	6.26	2	25	63	1	75	78	96
Compression0.5	96	10	0	12	6.50	5.19	5.00	4.55	6.82	3	36	94	3	79	64	92
Compression0.25	100	10	0	6	5.66	4.49	4.40	2.61	5.92	12	33	100	4	78	60	97
Centralization14	96	14	0	24	6.16	5.93	6.00	5.50	8.64	0	28	76	9	69	57	90
Centralization12	96	12	0	24	9.22	6.77	6.00	5.57	7.58	1	27	77	8	71	65	94
Centralization8	99	8	0	24	7.53	6.66	7.00	5.37	5.51	0	34	85	3	80	81	95
Centralization6	99	6	0	24	6.00	6.70	7.00	5.48	4.27	1	40	79	0	91	82	91
Average	98.38				7.71	6.63	6.00	5.43	6.44	3.46	31.54	80.85	3.62	79.85	72.92	94.62

Notes: Number of respondents in column 1 (obs). Columns 2 to 4 give information about the response scale: Number of intervals (#), center of the response scale, span of the closed intervals. Columns 5 to 7 give mean, trimmed mean (trimmed at 10 percent), and median response for the point forecast (Q3). Columns 8 and 9 show average mean forecasts (beta), and the average number of intervals used by the respondents. Columns 10 to 13 give information about response attitudes: The number of respondents using a single interval (10), using the full set of intervals (11) which is generally 10 but varies in the *Centralization* treatments, using any of the two open intervals (12), and providing responses with gaps (13). Columns 14 and 15 report the number of responses with uni-modal (single-peaked) response: uni-modal bar-chart of probabilities (14), uni-modal histogram (15). Column (16) reports the number of respondents expecting inflation in the binary inflation/deflation question.

3.D Attrition and Randomization checks for the US survey

Table A3.2 shows that in the US survey, respondents are equally likely to drop out of any of the treatments. Columns 1 and 2 show information on the number of all surveys started, while columns 3 and 4 contain information on all surveys that were actually finished by respondents. Columns 5 and 6 depict the number of complete surveys, that is all finished surveys minus those respondents i) that did not pass the attention check (1 respondent), ii) that were living outside the US (1 respondent), and iii) that provided answers to the density forecast that do not add up to 100 (20 respondents). Attrition overall was very low.

Table A3.2: Attrition check. Number of participants that started, finished, and completed the surveys (by treatment).

	All surveys		Finished surveys		Complete surveys	
Treatment	Participants	Share	Participants	Share	Participants	Share
Baseline	105	7.74	101	7.76	101	7.90
ShiftMinus12	102	7.52	100	7.69	99	7.74
ShiftMinus4	107	7.89	100	7.69	99	7.74
ShiftPlus4	106	7.82	100	7.69	98	7.66
ShiftPlus12	103	7.60	100	7.69	98	7.66
Compression4	103	7.60	100	7.69	99	7.74
Compression2	102	7.52	100	7.69	99	7.74
Compression0.5	103	7.60	100	7.69	96	7.51
Compression0.25	105	7.74	100	7.69	100	7.82
Centralization14	107	7.89	100	7.69	96	7.51
Centralization12	104	7.67	100	7.69	96	7.51
Centralization8	106	7.82	100	7.69	99	7.74
Centralization6	103	7.60	100	7.69	99	7.74
Sum	1356	100	1301	100	1279	100

Additionally we compare the demographics shown in Table A3.2 against data from the US Census Bureau from 2021. We compare the mean age against the census mean age, accounting for the fact that respondents on Prolific has to be of age 18 or older. We also compare the shares for respondents identifying as female, black, or white against the shares reported in US census data. We find that our overall sample is representative of the US population in terms of the share of female, black, and white respondents. However, or sample is on average around 2 years younger than the average US citizen (diff$=-2.06$, $p < 0.001$, t-test). In terms of the individual treatments, we find differences in age at the 5% level for ShiftPlus4, ShiftPlus12, Compression2, and Centralization12 , as well as the 10% level for Centralization6 (t-tests). For share of females, we find differences at the 5% level for Compression4 and Compression0.25, and at the 10% level for ShiftMinus4 (proportions z-tests). For white respondents, we find differences at the 5% level for Compression2 and Centralization8, and at the 10% level for ShiftPlus4 (proportions z-tests). The share of black respondents is only significant at the 5% level in Compression4 (proportions z-tests). Overall, the one treatment showing strong differences is Compression4, with a significantly higher share of females (0.61), significantly higher share of white (0.84), and significantly lower share of black respondents (0.05). Looking more closely at this treatment, this is caused by Compression4 having a substantially larger number of white female respondents (52 out of 99 respondents). Note that if we control for multiple hypotheses testing using the Bonferroni-Holm method, only the aforementioned age difference for the overall sample would remain (strongly) significant.

3.E Respondents' proficiency regressions

Tables A3.3 to A3.6 in this appendix report regressions of the probability mass respondents' assign to certain ranges of inflation on treatment dummies, interaction terms with respondents' proficiency, and other controls. A respondent's proficiency refers to one of the following three measures: financial literacy, highest obtained degree, or knowledge of the Fed's inflation target. Financial literacy was elicited in a questionnaire at the end of the experiment. We used the three-item financial literacy test by Lusardi & Mitchell (2014), *Financial lit.*, ranging from 0 to 3 correct answers. In addition, we asked respondents for their knowledge of the Federal Reserve Bank's inflation target (*Target correct* dummy) and their level of education (*Education high* dummy indicates a BA degree or higher).
As outlined in Sections 3.3.2 to 3.3.2, the different treatments require different test ranges: All *Shift* treatments are evaluated via the range of deflation, $(-\infty, 0]$; *Compression4, Compression2,* and *Centralization6* via the range $[-8, 8]$; *Compression0.5* and *Centralization8* via the range $[-4, 4]$; and *Compression0.25* and *Centralization14* via the range $[-2, 2]$.
Subjective answer certainty was elicited via a 6-item Likert scale (*Certain,* ranging from 0 = Very Uncertain to 5 = Very Certain) asked directly after each inflation expectation question. The regressions use age, gender, and political orientation. In the survey, we also elicited the state of residence from respondents and use this information to create region dummy variables based on the definition of the US Census Bureau (West, Midwest, South, Northeast, Territories). The regressions use these dummies to control for region of residence.

Table A3.7 shows the results of regressions of the responses on the three certainty questions on respondents' proficiency (i.e., *Financial lit.*, *Target correct*, and *Education high*), the corresponding forecast, as well as controls for age, gender, political orientation, and region. For specifications (1)-(3), the *Certain* variable on the left side of the regression equation refers to respondents' certainty answer after the density forecast, for specifications (4)-(6), it is the certainty answer after the binary inflation/deflation forecast, and for specifications (7)-(9), it is the certainty answer after the point prediction. *Forecast* is the respondent's forecasts preceding the certainty question: For specifications (1)-(3), it is the mean of the fitted beta distribution, a dummy for predicting inflation for (4)-(6), and the value of the point prediction for (7)-(9).

Table A3.3: Shift Treatments. OLS regressions of the probability mass assigned to deflation on respondents' proficiency and interactions.

Probability Mass in Deflation	(1)		(2)		(3)	
ShiftMinus12	26.36***	(0.000)	27.58***	(0.000)	30.06***	(0.005)
ShiftMinus4	9.047***	(0.007)	9.588***	(0.003)	28.93***	(0.003)
ShiftPlus4	-3.440	(0.301)	-5.024	(0.110)	-5.779	(0.586)
ShiftPlus12	-6.276*	(0.059)	-8.427***	(0.008)	-24.97***	(0.009)
Certain			-5.707***	(0.000)	-5.467***	(0.000)
Financial lit.					-5.148**	(0.046)
Financial lit. × ShiftMinus12					5.140	(0.245)
Financial lit. × ShiftMinus4					-6.029	(0.128)
Financial lit. × ShiftPlus4					1.498	(0.733)
Financial lit. × ShiftPlus12					6.980*	(0.055)
Target correct=1					-4.779	(0.281)
Target correct=1 × ShiftMinus12					-8.278	(0.203)
Target correct=1 × ShiftMinus4					6.139	(0.350)
Target correct=1 × ShiftPlus4					4.218	(0.507)
Target correct=1 × ShiftPlus12					2.120	(0.738)
Education high=1					0.0854	(0.985)
Education high=1 × ShiftMinus12					-15.22**	(0.022)
Education high=1 × ShiftMinus4					-11.08*	(0.092)
Education high=1 × ShiftPlus4					-7.049	(0.287)
Education high=1 × ShiftPlus12					-2.042	(0.752)
Constant	9.307***	(0.000)	30.35***	(0.000)	42.41***	(0.000)
Controls	No		Yes		Yes	
Observations	495		494		492	
Adjusted R^2	0.204		0.294		0.372	

Notes: p-values in parentheses. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Table A3.4: Compression4, Compression2, Centralization6 Treatments. OLS regressions of the probability mass assigned to the intervals in the range $[-8, 8]$ on respondents' proficiency and interactions.

Probability mass in range $[-8, 8]$	(1)		(2)		(3)	
Compression4	-15.75***	(0.000)	-12.54***	(0.003)	-41.97***	(0.000)
Compression2	-7.223*	(0.095)	-4.621	(0.270)	-16.40	(0.181)
Centralization6	-9.556**	(0.027)	-8.460**	(0.041)	-25.14*	(0.056)
Certain			-1.499	(0.217)	-1.773	(0.141)
Financial lit.					-2.078	(0.548)
Financial lit. × Compression4					13.74***	(0.004)
Financial lit. × Compression2					5.364	(0.270)
Financial lit. × Centralization6					8.766*	(0.088)
Target correct=1					5.070	(0.397)
Target correct=1 × Compression4					-5.534	(0.508)
Target correct=1 × Compression2					-12.56	(0.145)
Target correct=1 × Centralization6					-3.828	(0.646)
Education high=1					8.162	(0.192)
Education high=1 × Compression4					-1.379	(0.875)
Education high=1 × Compression2					10.21	(0.246)
Education high=1 × Centralization6					-3.316	(0.699)
Constant	68.46***	(0.000)	66.22***	(0.000)	63.88***	(0.000)
Controls	No		Yes		Yes	
Observations	398		398		398	
Adjusted R^2	0.026		0.113		0.161	

Notes: p-values in parentheses. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Table A3.5: Compression0.5, Centralization8 Treatments. OLS regressions of the probability mass assigned to the intervals in the range $[-4, 4]$ on respondents' proficiency and interactions.

Probability mass range $[-4, 4]$	(1)		(2)		(3)	
Compression0.5	1.598	(0.703)	3.324	(0.421)	-16.43	(0.212)
Centralization8	-7.920*	(0.058)	-6.462	(0.117)	-19.29	(0.121)
Certain			-4.498***	(0.004)	-4.997***	(0.001)
Financial lit.					-5.262	(0.134)
Financial lit. × Compression0.5					11.10**	(0.027)
Financial lit. × Centralization8					6.669	(0.188)
Target correct=1					-0.792	(0.895)
Target correct=1 × Compression0.5 1					-7.927	(0.352)
Target correct=1 × Centralization8					15.42*	(0.070)
Education high=1					3.975	(0.527)
Education high=1 × Compression0.5					-4.207	(0.629)
Education high=1 × Centralization8					-17.39*	(0.051)
Constant	34.48***	(0.000)	50.24***	(0.000)	59.92***	(0.000)
Controls	No		Yes		Yes	
Observations	296		296		295	
Adjusted R^2	0.013		0.058		0.080	

Notes: p-values in parentheses. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Table A3.6: Compression0.25, Centralization14, Centralization12 Treatments. OLS regressions of the probability mass assigned to the intervals in the range $[-2, 2]$ on respondents' proficiency and interactions.

Probability mass range $[-2, 2]$	(1)		(2)		(3)	
Compression0.25	13.04***	(0.000)	14.63***	(0.000)	18.86**	(0.044)
Centralization14	7.448**	(0.017)	6.915**	(0.017)	6.723	(0.471)
Centralization12	4.281	(0.168)	3.208	(0.269)	-2.418	(0.784)
Certain			-5.088***	(0.000)	-5.073***	(0.000)
Financial lit.					-4.371*	(0.072)
Financial lit. × Compression0.25					0.584	(0.877)
Financial lit. × Centralization14					1.909	(0.592)
Financial lit. × Centralization12					4.353	(0.214)
Target correct=1					2.467	(0.555)
Target correct=1 × Compression0.25					-14.17**	(0.016)
Target correct=1 × Centralization14					-10.01*	(0.099)
Target correct=1 × Centralization12					-5.742	(0.329)
Education high=1					-1.706	(0.696)
Education high=1 × Compression0.25					2.004	(0.736)
Education high=1 × Centralization14					2.255	(0.713)
Education high=1 × Centralization12					-3.307	(0.586)
Constant	13.00***	(0.000)	33.70***	(0.000)	42.71***	(0.000)
Controls	No		Yes		Yes	
Observations	393		392		392	
Adjusted R^2	0.040		0.183		0.204	

Notes: p-values in parentheses. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

Table A3.7: Certainty. OLS regressions of respondents' certainty on her proficiency, her corresponding forecast, and controls, for each of the three certainty questions.

Probability mass range $[-2, 2]$	(1)		(2)		(3)	
Compression0.25	13.04***	(0.000)	14.63***	(0.000)	18.86**	(0.044)
Centralization14	7.448**	(0.017)	6.915**	(0.017)	6.723	(0.471)
Centralization12	4.281	(0.168)	3.208	(0.269)	-2.418	(0.784)
Certain			-5.088***	(0.000)	-5.073***	(0.000)
Financial lit.					-4.371*	(0.072)
Financial lit. × Compression0.25					0.584	(0.877)
Financial lit. × Centralization14					1.909	(0.592)
Financial lit. × Centralization12					4.353	(0.214)
Target correct=1					2.467	(0.555)
Target correct=1 × Compression0.25					-14.17**	(0.016)
Target correct=1 × Centralization14					-10.01*	(0.099)
Target correct=1 × Centralization12					-5.742	(0.329)
Education high=1					-1.706	(0.696)
Education high=1 × Compression0.25					2.004	(0.736)
Education high=1 × Centralization14					2.255	(0.713)
Education high=1 × Centralization12					-3.307	(0.586)
Constant	13.00***	(0.000)	33.70***	(0.000)	42.71***	(0.000)
Controls	No		Yes		Yes	
Observations	393		392		392	
Adjusted R^2	0.040		0.183		0.204	

Notes: p-values in parentheses. */**/*** denotes significance at the 0.1/0.05/0.01 probability level.

3.F Mechanical treatment effects

Modifying the response scale may affect mean and uncertainty in Table A3.8 independently of any behavioral biases.[46] Spurious treatment effects may show up when a large part of the probability mass is assigned to an open interval or when the intervals are "too wide". As an example of the latter case, imagine a respondent who expects inflation to fall into the narrow range from 0 to 2 percent. The response scale in *Baseline* has narrow intervals in this range and thus allows the respondent to provide a histogram that closely reflect her beliefs. But other response scales, such as *Compression4* where the corresponding interval is from 0 to 8 percent, would distort the respondent's beliefs and we would overestimate the respondent's mean and uncertainty.

Table A3.8: Mechanical treatment effects.

Distribution	$\mathcal{N}(0,4)$		$\mathcal{N}(0,9)$		$\mathcal{N}(4,4)$		$\mathcal{N}(4,9)$		$\mathcal{N}(8,4)$		$\mathcal{N}(8,9)$	
	Mean	Uncertainty	Mean	Uncertainty	Mean	Uncertainty	Mean	Uncertainty	Mean	Uncertainty	Mean	Uncertainty
Theoretical	0.00	2.00	0.00	3.00	4.00	2.00	4.00	3.00	8.00	2.00	8.00	3.00
Baseline	0.00	2.18	0.00	3.24	4.23	2.20	4.14	3.17	8.07	2.42	8.22	3.47
ShiftMinus12	0.91	3.15	0.62	3.57	3.86	0.89	3.44	1.79	4.00	0.03	3.98	0.37
ShiftMinus4	0.23	2.20	0.14	3.17	4.07	2.42	4.22	3.47	8.91	3.15	8.62	3.57
ShiftPlus4	-0.23	2.20	-0.14	3.17	4.00	2.18	4.00	3.24	8.23	2.20	8.14	3.17
ShiftPlus12	-0.91	3.15	-0.62	3.57	3.93	2.42	3.78	3.47	7.77	2.20	7.86	3.17
Compression4	0.00	4.00	0.00	4.12	4.00	1.71	4.00	3.42	8.00	4.00	8.02	4.17
Compression2	0.00	2.34	0.00	3.26	4.05	2.47	4.17	3.55	8.91	3.15	8.65	3.64
Compression0.5	0.00	2.11	0.00	3.14	4.14	2.21	4.04	2.97	7.48	1.24	7.01	1.84
Compression0.25	0.00	2.02	0.00	2.62	3.28	1.21	2.86	1.78	3.99	0.14	3.89	0.54
Centralization14	0.00	2.21	0.00	3.25	4.26	2.16	4.17	3.15	8.07	2.42	8.22	3.46
Centralization12	0.00	2.20	0.00	3.25	4.26	2.16	4.17	3.15	8.07	2.42	8.22	3.46
Centralization8	0.00	2.34	0.00	3.22	4.00	2.34	4.01	3.24	8.05	2.47	8.17	3.55
Centralization6	0.00	4.00	0.00	4.08	3.95	1.51	3.84	3.08	7.14	3.26	7.52	3.97

Notes: The table shows mean and uncertainty under the assumption of normally distributed inflation expectations $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2. Mean and uncertainty calculated using a mass-at-midpoint assumption after binning the normally distributed data in the intervals of the response scales.

To illustrate the magnitude of these effects, consider a hypothetical setting in which a household with fixed probabilistic expectations is confronted with the response scales of the 13 treatments. Assuming that the household's expectations are normally distributed, we calculate the probability mass assigned to each interval and compute mean and uncertainty using a simple mass-at-midpoint measure (the results are similar when we follow Engelberg et al. (2009) and Becker et al. (2022) and calculate a smoothed response instead). For the household's normally distributed beliefs, we assume means of 0, 4, and 8 and variances of 4 and 9 to capture settings with low and high inflation uncertainty. Table A3.8 presents the results. Regarding the mean, the mechanical treatment effects are typically small, except in cases where a large part of the probability mass is assigned to an open interval (i.e., *ShiftMinus12* and *Compression0.25*). Regarding uncertainty, we observe mechanical treatment effects when the open intervals contain a large part of the probability mass and when the intervals are comparatively wide (e.g., *Compression4*).

[46] In order to avoid these "mechanical treatment effects" the tests in Sections 3.3.2 to 3.3.2 compare the probability masses the respondents assign to specific ranges of inflation. The mechanical treatment effects we describe in this appendix are absent in these probability-mass-tests.

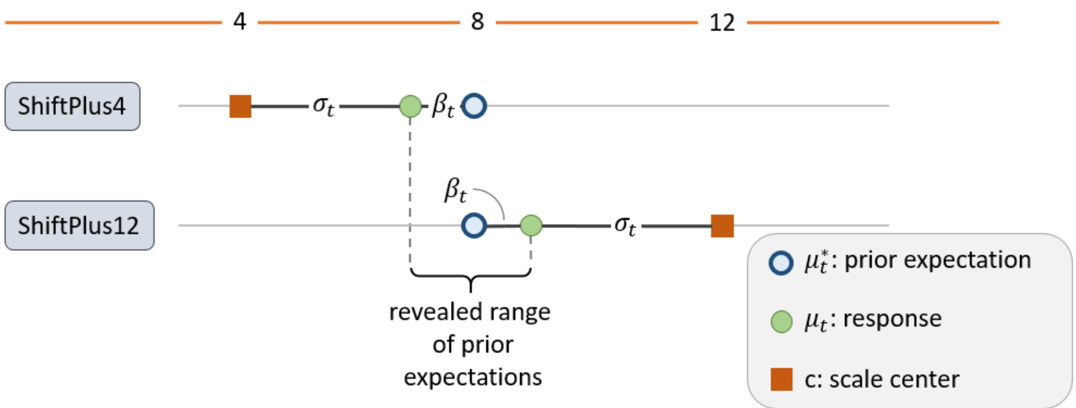
3.G Bounds on respondents' prior expectations

Under a comparatively strong assumption, the *Shift* treatments allow the identification of bounds on respondents' prior (or true) expectations. The prior expectations are a respondent's expectations before observing the response scale. This appendix sketches the idea. The central assumption is that the average mean response, denoted by μ_t, lies between the center of the scale, c, and respondents' prior expectations μ_t^*

$$c \leq \mu_t \leq \mu_t^* \quad \text{or} \quad c \geq \mu_t \geq \mu_t^*.$$

Figure A3.4 illustrates the scale center, the average mean response, and a possible location of the respondents' average prior distribution for *ShiftPlus4* and *ShiftPlus12*. $\beta_t = (\mu_t^* - \mu_t)$ denotes the average measurement bias and $\sigma_t = (\mu_t - c)$ the distance between the average mean response and the scale center. Using the numbers from Table 3.2, the average mean response in *ShiftPlus12* is 8.34 which is lower than the center of the response scale (12). Under our assumption, we may then conclude that 8.34 is an upper bound for μ_t^*. In *ShiftPlus4*, the average mean response (6.83) is larger than the center of the response scale (4) and we may conclude that 6.83 is a lower bound for μ_t^*. Notably, both the median and the trimmed mean of point forecast fall within these bounds, see Table A3.1.

Figure A3.4: Constructing bounds on respondents' prior expectations μ_t^* under the assumption that the average mean response μ_t lies between the center of the response scale and respondents' prior expectations μ_t^*.



How plausible is the assumption that the average mean response lies between the center of the scale and respondents' prior expectations? Respondents' tendency to move their answers towards the center of the scale (Result 1) is strong and it seems plausible that the majority of responses follow this pattern. However, violations of this assumption are possible. For example, even if we assume that such an ordering holds for every individual, it may be violated in the aggregate. In addition, the mechanical effects described in Appendix F may influence individual $\mu_{i,t}$ and therefore the ordering. Other violations are conceivable. We leave a rigorous treatment of this idea for future research.

Chapter 4

Households' Probabilistic Inflation Expectations in High-Inflation Regimes

Abstract¶

Central bank surveys frequently elicit households' probabilistic beliefs about future inflation. The responses provide only a coarse picture of inflation beliefs further away from zero. Using data from the Bundesbank household panel, we show that the current high-inflation environment induces respondents to allocate considerable probability to the rightmost open interval. This pile-up of probabilities negatively affects estimates of histogram moments and leads to a divergence between average expected inflation measured by probabilistic and point forecasts. The consistency of predictions can be improved by using an alternative design of the response scale that allows respondents to state more detailed beliefs for higher inflation ranges.

¶Joint work with Peter Duertsch, Thomas Eife and Alexander Glas.

We are grateful to the Bundesbank for including our questions in the Bundesbank Online Panel Households. Our research was improved by very helpful comments and suggestions of Jonas Dovern, Johannes Frank, Norbert Schwarz and Michael Weber.

4.1 Introduction

Survey data are a popular source of information about the macroeconomic expectations of experts, households and firms. In addition to point forecasts, many surveys provide probabilistic expectations which are typically elicited by asking respondents to assign probabilities to predefined outcome intervals ("bins"). These probability distributions offer important insights into how survey participants assess the uncertainty, skewness and tail risk associated with their predictions (Manski, n.d.).

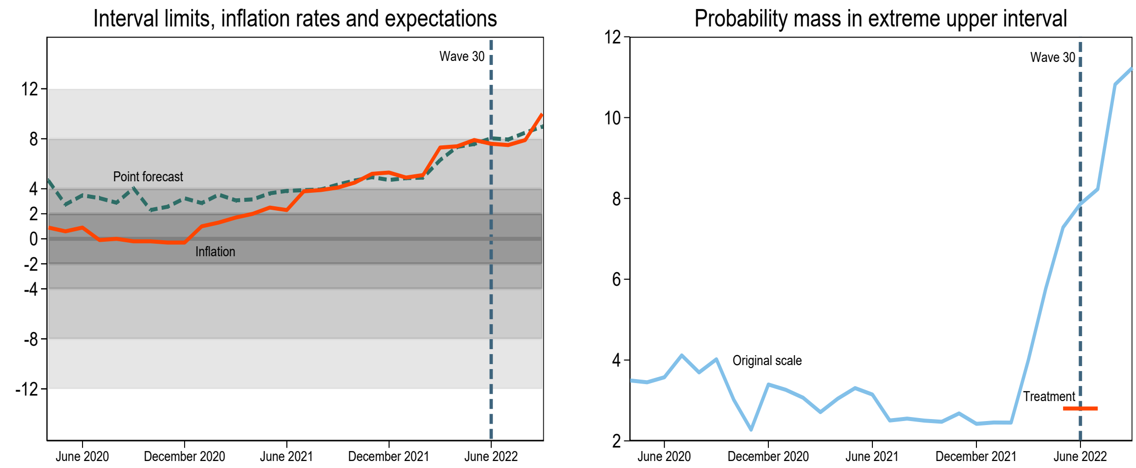
In this paper, we analyze the quality of the probabilistic inflation expectations measured in the Bundesbank Online Panel Households (BOP-HH) in light of the recent surge of inflation in Germany and the euro area as a whole. In particular, we assess whether adjusting the bin definitions improves the consistency between the point forecasts and the probabilistic expectations by conducting a randomized experiment where some of the participants in Wave 30 (June 2022) receive the original bin design, while others receive an alternative design where the center of the intervals is closer to—but still below—the actual German inflation rate.

Our central finding is that the alternative design leads to considerably more consistent responses with the probabilistic expectations closely matching both actual inflation and point forecasts. This improved match between point forecasts and probabilistic expectations is driven by the fact that the original scale offers respondents a relatively small set of reasonable choices at times when inflation is very low or very high. For example, respondents who expect inflation rates of eight percent or higher only have two intervals at their disposal. This forces them either to provide inconsistent answers or to assign probabilities in extreme, marginal intervals, which is something that many respondents tend to avoid (Becker et al., 2023). Our finding is relevant for all surveys that employ scales similar to the one used in BOP-HH.[47]

As illustrated in the left plot in Figure 4.1, the question about probabilistic expectations consists of ten bins which are centered around an inflation rate of 0%. The interior bins cover the range from −12% to +12%. The two exterior bins are half-open. A major advantage of using this responce scale is that it allows for a comparison of results both within surveys (across time) and between surveys (across different geographical locations). The red line shows the monthly German inflation rate based on the consumer price index. Before 2021, inflation rates were close to the center of the response scale. Inflation began to rise during the COVID-19 pandemic and further accelerated after the Russian invasion of Ukraine in February 2022 and the associated energy crisis. The inflation rate in June 2022, when our experiment was conducted, was 7.9%, which is just slightly below the lower bound of the rightmost interior bin. By September 2022, inflation further increased to 10%. The green line shows average inflation expectations in the BOP-HH. Clearly, households take notice of this development and adjust their point forecast accordingly.

[47] The baseline definition used in the BOP-HH was originally designed for the Federal Reserve Bank of New York's Survey of Consumer Expectations. See Armantier et al. (2017) for an overview. Other examples include the European Central Bank's Consumer Expectations Survey and similar surveys conducted by the central banks of Canada, France, the Netherlands, Ukraine, and the United Kingdom.

Figure 4.1: Probabilistic inflation expectations and interval definitions.



Notes: The left plot shows monthly German consumer price inflation (red line). The dashed green line depicts the average inflation expectations of German households (trimmed by 1% from bottom and top in each month). The shaded gray areas correspond to the original bin definitions in the BOP-HH. The dashed blue line indicates the June 2022 wave of the BOP-HH to which we contributed an alternative bin design. The right plot shows the average probability mass in the highest bin based on participants presented with the original bin design (blue line). The red bar shows the corresponding average probability mass for the individuals presented our alternative bin design.

The increase in households' point predictions is accompanied by an upward shift in their probabilistic inflation expectations. The blue line in the right plot shows the average probability mass assigned to the rightmost (half-open) bin. Before February 2022, the average probability fluctuated at relatively low levels between 2% and 4%. Consistent with the higher average point forecasts, we observe a steep increase in the average probability since the Russian invasion of Ukraine. The average probability in the rightmost bin was 7.9% in June 2022 and rose even further to more than 11% by September 2022. Since it is unknown what respondents consider a likely upper bound for inflation, the information provided by the open interval is limited. One has to make an assumption about the upper bound to derive a belief distribution from the answers. Thus, the evidence in Figure 4.1 puts into question the reliability of moments derived from the probabilistic expectations based on the original survey design.

We contributed an alternative bin design to Wave 30 of the BOP-HH where the center of the intervals is shifted from 0% to 4%, while keeping the relative bin width identical to the original design. As a result, the interior bins in the alternative treatment cover a range from -8% to $+16\%$. The red bar in the right plot shows that for this treatment group, the average probability mass assigned to the rightmost bin is 2.8%, which is much more in line with the figures observed in earlier survey waves. These respondents use more bins, report higher histogram means that are more consistent with their point forecasts and report lower uncertainty than those in the baseline group. We conclude that the distortion of moments of the obtained belief distribution can be reduced by adjusting the bin definitions at times when inflation is unusually high.

Our research relates to the literature that explores how households form their macroeconomic expectations. Important covariates include households' socioeconomic characteristics such as gender, income and education (Bruine de Bruin et al., 2010; Das et al., 2020), their sources of information about monetary policy and the state of the economy (Coibion et al., 2022; Conrad et al., 2022) as well as individual and macroeconomic lifetime experiences (Malmendier & Nagel, 2011, 2016; D'Acunto, Malmendier, et al., 2021). Using the BOP-HH data, Conrad et al. (2022) show that households' quantitative inflation expectations are related to the information channels that households use to inform themselves about monetary policy. In contrast, their qualitative expectations, i.e., the expected future direction of inflation, is more closely related to an individuals' lifetime inflation experiences. While these studies focus on households' point forecasts, we consider probabilistic expectations. Using the Michigan Survey of Consumers, Bruine de Bruin et al. (2011) show that consumers are generally willing and able to provide meaningful probability distributions that are consistent with the point predictions. Similarly, Zhao (2022) finds that the point forecasts of US households in the Federal Reserve Bank of New York's Survey of Consumer Expectations tend to be well-aligned with their probabilistic expectations. We contribute to the literature by analyzing whether the quality of the probabilistic expectations is related to the formulation of the corresponding question in the survey questionnaire in high-inflation regimes. As such, our analysis also relates to the literature that analyze how specifics of the survey design influence the responses. Here, Schwarz (2010) gives a good overview in general while Becker et al. (2023) and Weber et al. (2022) discuss this point in the context of inflation expectations.

The rest of this paper is organized as follows. Section 4.2 explains the data and discusses the competing designs of the question used for the probabilistic inflation expectations. Section 4.3 presents the results. We discuss our findings in Section 4.4. Section 4.5 concludes.

4.2 Bundesbank Online Panel Households

We use data from the BOP-HH, a representative online survey of German households operated by the Bundesbank. The survey targets individuals aged 16 years or older (see Beckmann & Schmidt, 2020, for details on the elicitation process). Among other questions, participants are asked to state their inflation expectations and socioeconomic characteristics. The survey started in 2019 with three pilot surveys. Starting with Wave 4 (April 2020), the BOP-HH is issued on a monthly basis. We focus on the responses from Wave 30 (June 2022) to which we contributed alternative formulations for the question on the probabilistic inflation expectations. In Section 4.3.4, we consider revisions of inflation expectations by comparing the responses from Wave 30 to those in Wave 29 (May 2022) and 31 (July 2022).

In total, 4,460 households participated in Wave 30. We remove observations from the sample whenever the household did not report probabilistic inflation expectations or if information for any of the socioeconomic characteristics is missing. We also exclude one respondent who did not state whether her point forecast represents a deflation rate or an inflation rate. This leaves 4,094 observations in our sample for Wave 30.

4.2.1 Probabilistic inflation expectations

BOP-HH participants receive the following question on their probabilistic expectations:[48]

`CM004: In your opinion, how likely is it that the rate of inflation will change as follows over the next twelve months?`

- `The rate of deflation (opposite of inflation) will be 12% or higher.`
- `The rate of deflation ([...]) will be between 8% and less than 12%.`
- `The rate of deflation ([...]) will be between 4% and less than 8%.`
- `The rate of deflation ([...]) will be between 2% and less than 4%.`
- `The rate of deflation ([...]) will be between 0% and less than 2%.`
- `The rate of inflation will be between 0% and less than 2%.`
- `The rate of inflation will be between 2% and less than 4%.`
- `The rate of inflation will be between 4% and less than 8%.`
- `The rate of inflation will be between 8% and less than 12%.`
- `The rate of inflation will be 12% or higher.`

Respondents are asked to rate the probability of inflation falling into each bin on a scale from 0 to 100, with 0 meaning that this outcome is completely unlikely and 100 meaning that they are absolutely certain it will happen. They also receive a notification that probabilities should add up to 100%. As mentioned above, the ten bins are centered around an inflation rate of 0%. Motivated by the recent surge in inflation rates, we contributed the following alternative bin design to the questionnaire of Wave 30:

`P3001A: In your opinion, how likely is it that the rate of inflation will change as follows over the next twelve months?`

- `The rate of deflation (opposite of inflation) will be 8% or higher.`
- `The rate of deflation ([...]) will be between 4% and less than 8%.`
- `The rate of deflation ([...]) will be between 0% and less than 4%.`
- `The rate of inflation will be between 0% and less than 2%.`
- `The rate of inflation will be between 2% and less than 4%.`
- `The rate of inflation will be between 4% and less than 6%.`
- `The rate of inflation will be between 6% and less than 8%.`
- `The rate of inflation will be between 8% and less than 12%.`
- `The rate of inflation will be between 12% and less than 16%.`
- `The rate of inflation will be 16% or higher.`

In the new formulation, the center of the bins is shifted upwards by four percentage points. As a result, the bins are centered around an inflation rate of 4%, which is closer to—but still below—the actual inflation rate in May 2022 (7.9%) relative to the baseline design. We leave

[48] All questions related to inflation include an info box that informs respondents that inflation is defined as the percentage change in the general price level as measured by the consumer prices index. They also receive the information that deflation is the opposite of inflation.

the number of bins as well as their widths unchanged.[49]

The sample in Wave 30 was split into three randomly assigned groups. Approximately one third of the sample (1,356 observations) was presented with the baseline design used in all previous waves. Another third of the sample (1,377 observations) was presented with the alternative design which we refer to as the 'mean-shift' setting. The remaining 1,361 observations were presented with another bin design which we do not use in our analysis.[50] Thus, our analysis focuses on the 2,733 households in the baseline group and the mean-shift group.

In the analysis below, we analyze the impact of the alternative response scale on households' probabilisitic expectations. We are particularly interested in potential differences in the shape of the histograms between the baseline group and the mean-shift group. Figure 4.2 shows the average responses of the individuals in both subsamples. The plot on the left depicts the average probability mass assigned to each bin across all respondents while the plot on the right shows the corresponding histogram by reporting densities instead of probabilities. The aggregate distributions clearly differ across treatments.

To assess the differences in the probabilistic expectations on an individual basis, we define the dummy variable *meanshift* that equals one if the individual belongs to the mean-shift group, and zero else. Next, we calculate the number of bins with nonzero probability (*bins*) and the probability mass assigned to the rightmost bin (*phigh*). We also define the dummy variable *multipeak* which equals one if the histogram has multiple modes, and zero else. Table A4.1 in the appendix provides details on the construction of all variables.

Finally, we compute the first four moments of each histogram. To do so, we follow Conrad et al. (2022) and assume that the probability in each bin is located at the midpoint.[51] To close the exterior bins, we assume that they have equal width to the adjacent bins, i.e., four percentage points.[52] Based on the 'mass-at-midpoint' approach, mean (μ), standard deviation

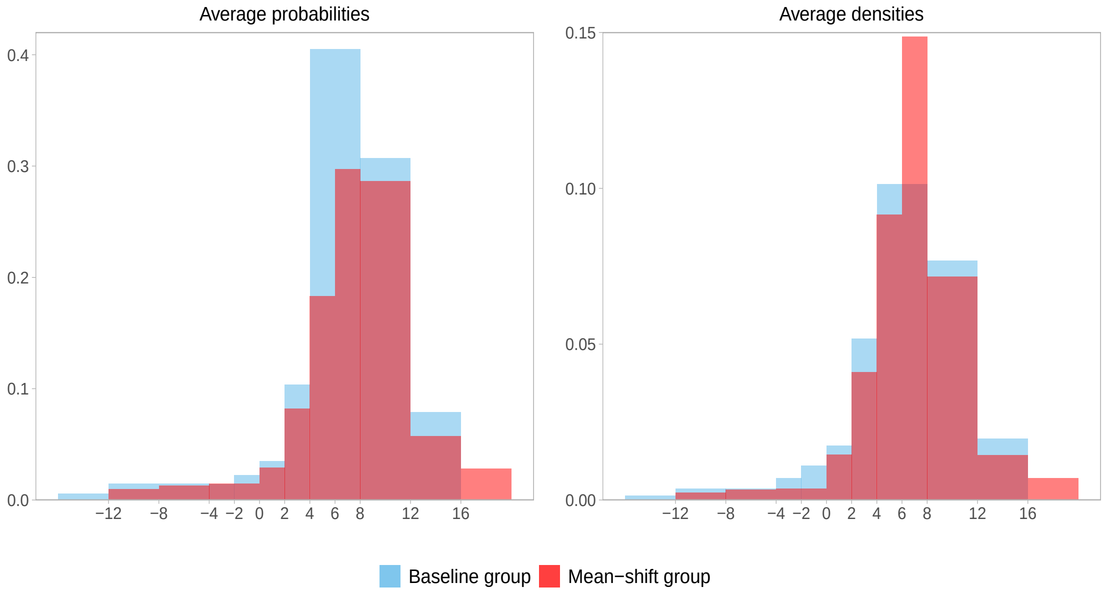
[49]Expert surveys such as the Survey of Professional Forecasters operated by the ECB and the Federal Reserve Bank of Philadelphia cover a relatively narrow outcome range. As a result, their operators frequently adjust the bin definitions in a way similar to our proposed mean-shift desgin. This is usually done in reponse to macroeconomic shocks such as the Great Recession where a considerable pile-up of probabilities in the lowest bin for GDP growth was observed in the ECB-SPF. During the Coronavirus pandemic, the ECB-SPF introduced bins with unequal width. Glas & Hartmann (2022) show that this can have an impact on the mismatch between ex-ante uncertainty as measured by the histogram variance and ex-post uncertainty based on the variability of forecast errors.

[50]This design retains the centering around 0% but includes a more granular definition of the interior bins. See Becker et al. (2023) for the motivation behind this approach. Since a takeaway from our study is that centering around an inflation rate of 0% is not appropriate in the current high-inflation regime, we do not use these observations in our analysis. However, all tables and figures for this alternative treatment are available upon request from the authors.

[51]Other alternatives include assuming uniformly distributed probabilities or fitting a continuous distribution as in Engelberg et al. (2009). However, Glas (2020) shows that this choice has little impact on estimates of the mean or the standard deviation. Moreover, Becker et al. (2022) show that fitting continuous distributions can lead to misleading results in the presence of varying interval widths.

[52]Armantier et al. (2017) and Zhao (2022) use ±38% as the bounds for the exterior bins in their analyses of the Survey of Consumer Expectations. This choice is based on historically observed inflation rates in the US. For Germany, such extreme inflation rates have not been observed. Zhao (2022) mentions in his footnote 14 that he also considered ±16% for the bounds and that this choice did not affect his findings. Our choice for the bounds also makes it more difficult to detect potentially existing differences between histogram means across treatments and when comparing histogram means to point forecasts.

Figure 4.2: Average probabilistic expectations by treatment status.



Notes: The subfigures show the average responses for the individuals in the baseline group and the mean-shift group. The left plot depicts the average probability mass in each bin while the right plot shows the histograms by reporting densities instead of probabilities.

(σ), skewness (γ) and kurtosis (κ) of the histogram reported by household $i = 1, \ldots, n$ are calculated as follows:

$$\mu_i = \sum_{k=1}^{K} m_k \times p_{i,k} \tag{4.1}$$

$$\sigma_i = \sqrt{\sum_{k=1}^{K} (m_k - \mu_i)^2 \times p_{i,k}} \tag{4.2}$$

$$\gamma_i = \frac{\sum_{k=1}^{K} (m_k - \mu_i)^3 \times p_{i,k}}{\sigma_i^3} \tag{4.3}$$

$$\kappa_i = \frac{\sum_{k=1}^{K} (m_k - \mu_i)^4 \times p_{i,k}}{\sigma_i^4} \tag{4.4}$$

In Equations (4.1)-(4.4), the index $k = 1, \ldots, K$ denotes the different bins, m_k is the midpoint of the k-th bin and $p_{i,k}$ is the probability assigned to this particular bin by household i.

Panel A of Table 4.1 presents summary statistics for all histogram characteristics by treatment status. For skewness and kurtosis, we consider only the responses of participants who use at least three bins. On average, the individuals in the mean-shift group use more bins, assign lower probability to the right-most bin, report higher histogram means and lower standard deviations.

Table 4.1: Summary statistics for Wave 30 of the BOP-HH.

	Baseline group					Mean-shift group				
	Obs.	Mean	SD	Min	Max	Obs.	Mean	SD	Min	Max
Panel A: Probabilistic inflation expectations										
bins	1,356	2.97	1.96	1.00	10.00	1,377	3.27	2.21	1.00	10.00
phigh	1,356	7.86	19.90	0.00	100.00	1,377	2.80	12.77	0.00	100.00
multipeak	1,356	0.08	0.26	0.00	1.00	1,377	0.08	0.27	0.00	1.00
μ_i	1,356	6.57	3.82	−14.00	14.00	1,377	7.24	3.41	−10.00	18.00
σ_i	1,356	2.02	1.74	0.00	11.34	1,377	1.82	1.74	0.00	11.72
γ_i	695	0.07	0.84	−4.14	5.02	778	0.11	0.82	−3.77	4.14
κ_i	695	3.61	2.44	1.22	29.62	778	3.53	2.24	1.08	25.73
Panel B: Point forecasts										
$\hat{\pi}_i^P$	443	6.63	2.53	0.00	20.00	435	6.67	2.55	0.00	20.00
$\hat{\pi}_i^E$	1,328	8.11	3.53	-2.00	30.00	1,350	8.14	3.32	-2.00	30.00
$\lvert\hat{\pi}_i^E - \mu_i\rvert$	1,328	2.17	3.40	0.00	25.40	1,350	1.60	2.54	0.00	26.35
Panel C: Socioeconomic characteristics										
age	1,356	56.98	14.35	17.00	80.00	1,377	56.83	14.62	16.00	80.00
east	1,356	0.17	0.38	0.00	1.00	1,377	0.17	0.37	0.00	1.00
female	1,356	0.36	0.48	0.00	1.00	1,377	0.39	0.49	0.00	1.00
fullemploy	1,356	0.44	0.50	0.00	1.00	1,377	0.42	0.49	0.00	1.00
hhsize	1,356	2.20	1.04	1.00	6.00	1,377	2.20	1.07	1.00	6.00
income	1,356	3.98	2.01	0.25	11.00	1,377	3.94	1.95	0.25	11.00
yoe	1,356	11.55	1.67	7.00	18.00	1,377	11.51	1.69	7.00	18.00

Notes: This table shows summary statistics for the probabilistic inflation expectations (Panel A), point forecasts (Panel B) and socioeconomic characteristics (Panel C) of participants in Wave 30 of the BOP-HH. For skewness and kurtosis, we focus on responses where nonzero probability is assigned to at least three bins. The samples for $\hat{\pi}_i^P$ and $\hat{\pi}_i^E$ are trimmed by 1% from top and bottom. Household income is expressed in 1,000 euro.

4.2.2 Point forecasts

In addition to the probabilistic expectations, the BOP-HH elicits point forecasts on households' perceptions of current inflation ($\hat{\pi}_i^P$) and their expectations of inflation over the coming year ($\hat{\pi}_i^E$). In the next section, we analyze the consistency of point and probabilistic expectations via the difference between $\hat{\pi}_i^E$ and μ_i. Since it has been shown that there exists a tight link between perceived and expected inflation (Jonung, 1981; D'Acunto, Hoang, et al., 2021), we also consider $\hat{\pi}_i^P$, although only one third of the participants in Wave 30 were asked for their perception of the current inflation rate over the previous twelve months. To reduce the impact of outliers, we trim the top and bottom 1% of inflation perceptions/expectations. For the remaining individuals, Figure 4.1 above shows average inflation expectations across survey waves along with actual inflation.

Panel B of Table 4.1 presents summary statistics for the point forecasts by treatment status. In contrast to the probabilistic expectations, the figures for perceived and expected inflation are very similar across the two treatment groups. Notably, the average point forecast exceeds the average histogram mean in both cases. However, due to the higher average histogram mean for the mean-shift group relative to the baseline group, the average absolute deviation between point forecasts and histogram means is markedly lower for this particular group.

The average perceived inflation rate (calculated as the weighted average across the two groups) is 6.65%. For comparison, the most recent inflation figure available to Wave 30 participants was the German inflation rate in May 2022 (7.9%) since all responses were collected between 15 June and 29 June and the May 2022 inflation rate was released by the German statistical office on 14 June. Only one response was elicited on 29 June when the first estimate of the inflation rate in June was released (7.6%). Thus, the average participant in Wave 30 *underestimates* current inflation. This finding contrasts the evidence in Conrad et al. (2022) who find that BOP-HH participants in Wave 3 overestimated inflation in May 2019. Their results are consistent with our data before June 2021 (see Figure 4.1). Weber et al. (2022) list a 'systematic upward bias' as a stylized fact of households' inflation perceptions/expectations. Our finding suggests that this may not generally be the case in high-inflation regimes or that households are slow to adjust their beliefs. However, the weighted average of expected inflation is 8.12%. Thus, households appear to take notice of the surge in inflation rates. This is supported by the upward trend in inflation expectations shown in Figure 4.1. The correlation between perceived and expected inflation is 0.46.

4.2.3 Socioeconomic characteristics

In addition to households' inflation expectations, we use information about their socioeconomic status. We consider age (*age*), gender (*female*), employment status (*fullemploy*), whether the individual lives in East or West Germany (*east*), household size (*hhsize*), income (*income*) and years of education (*yoe*). These variables have been shown to be robust predictors of households' macroeconomic expectations and uncertainty thereof (Bruine de Bruin et al., 2010, 2011; Das et al., 2020). In all regressions below, we use the natural logarithm of income as a covariate. We include these characteristics to improve the efficiency of the estimates.

Panel C of Table 4.1 presents summary statistics for the socioeconomic characteristics of the participants in Wave 30 by treatment status. The average respondent in Wave 30 is 57 years old and has almost 12 years of education. 38% of the individuals are female, 43% are full-time employed and 17% live in East Germany.

As with the point forecasts, socioeconomic characteristics are distributed similarly in both treatment groups, suggesting that the treatment is indeed randomly assigned. We confirm that this is the case by running a linear regression of the *meanshift*-dummy on the socioeconomic variables. The baseline is the group of households that were presented with the original bin design. Table A4.2 in the appendix presents the results. As expected, none of the coefficients are significantly different from zero, which suggests that the random assignment of treatments was successful.

Table 4.2: Inflation expectations and socioeconomic characteristics: baseline group.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
age	-0.03***	-0.08*	0.00	0.02**	-0.02***	-0.00	0.02***	0.01	-0.01
	(0.00)	(0.04)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
east	-0.10	4.10**	-0.01	0.76***	-0.11	0.01	0.12	0.73***	-0.09
	(0.15)	(1.74)	(0.02)	(0.27)	(0.13)	(0.09)	(0.36)	(0.26)	(0.23)
female	-0.16	4.88***	0.06***	0.50**	0.12	-0.17**	0.03	0.93***	0.60***
	(0.11)	(1.24)	(0.02)	(0.24)	(0.11)	(0.07)	(0.20)	(0.22)	(0.22)
fullemploy	-0.27**	0.57	-0.01	0.28	-0.28**	0.03	0.13	0.51**	0.24
	(0.13)	(1.30)	(0.02)	(0.25)	(0.11)	(0.07)	(0.21)	(0.25)	(0.24)
hhsize	0.02	0.97	-0.01	0.16	-0.03	-0.07**	-0.02	0.18	0.07
	(0.06)	(0.71)	(0.01)	(0.13)	(0.05)	(0.03)	(0.09)	(0.13)	(0.12)
ln(*income*)	0.07	-2.59**	-0.01	-0.45*	0.03	0.09	0.08	-0.89***	-0.41*
	(0.11)	(1.30)	(0.01)	(0.25)	(0.09)	(0.07)	(0.21)	(0.22)	(0.22)
yoe	0.01	-0.86**	-0.01***	-0.06	-0.04	0.06***	0.10*	-0.14**	-0.10*
	(0.03)	(0.40)	(0.00)	(0.07)	(0.03)	(0.02)	(0.06)	(0.06)	(0.06)
Constant	4.05***	38.69***	0.26**	9.11***	3.46***	-1.12**	0.60	15.50***	6.59***
	(0.93)	(9.65)	(0.11)	(2.14)	(0.82)	(0.56)	(1.63)	(1.67)	(1.80)
Observations	1,356	1,356	1,356	1,356	1,356	695	695	1,328	1,328
$\bar{R}^2$	0.03	0.03	0.02	0.01	0.02	0.02	0.01	0.04	0.01

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on socioeconomic characteristics. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

4.3 Results

This section presents our empirical findings. We briefly consider the relationship between inflation expectations and socioeconomic status before analyzing in which aspects the inflation expectations differ between the baseline and the mean-shift group. Next, we assess the implications of our results for the consistency between point forecasts and histogram means and explore potential heterogeneity in the estimated treatment effects. Lastly, we consider revisions in inflation expectations from Wave 29 (before the treatment) to Wave 30 as well as revisions from Wave 30 to Wave 31 (after the treatment).

4.3.1 Histogram characteristics and socioeconomic status

In a first step, we relate the histogram characteristics and point forecasts of BOP-HH participants to their socioeconomic status. Table 4.2 presents the estimates for the individuals in the baseline group. Columns (1)-(3) show the results for the number of bins with nonzero probability, the probability mass in the rightmost bin and the indicator for multimodal histograms. Columns (4)-(7) present the estimates for the histogram moments. Columns (8)-(9) show the findings for the point forecasts and the absolute deviations between point forecasts and histgram means. All regressions are estimated with heteroskedasticity-consistent standard errors.

Consistent with Armantier et al. (2021), we find that older respondents have significantly higher histogram means and lower inflation uncertainty as a result of using fewer bins. In addition, kurtosis increases with age. The *east*-dummy has a significantly positive effect on histogram means and point forecasts. This is in line with Goldfayn-Frank & Wohlfart (2020) who show that East Germans have higher inflation expectations than West Germans—especially at times when inflation is unusually high—due to the inflationary shock experienced after reunification. Next, we find that women assign more probability mass to the rightmost bin and have higher inflation expectations both in terms of histogram means and point forecasts. These findings square with similar evidence in Bruine de Bruin et al. (2011), Armantier et al. (2021) and Conrad et al. (2022). In addition, the probability of reporting a multi-peaked probability distribution is significantly higher for women, the histograms of women are more left-skewed than those of men and their point forecasts and histogram means tend to deviate more strongly. Full-time employed individuals use fewer bins, have lower uncertainty and higher point forecasts (but not histogram means). Household size appears to matter little beyond a negative effect on skewness. Higher income is associated with a lower probability mass in the rightmost bin (as in Armantier et al., 2021), lower point forecasts and a higher degree of consistency between point forecasts and histogram means (see Zhao, 2022). Lastly, higher education is associated with less probability in the rightmost bin, a lower probability of stating a multi-peaked distribution, higher skewness and kurtosis, lower point forecasts and smaller deviations between point forecasts and histogram means. The findings that high-income households and highly educated individuals have lower point forecasts are consistent with Bruine de Bruin et al. (2010).

Overall, our results are in line with typical findings in the literature (Das et al., 2020). In the following analyses, we use each respondents' socioeconomic characteristics as control variables in all regressions. Since treatment assignment is unrelated to socioeconomic characteristics (see Table A4.2), these variables are included primarily to increase the efficiency of the estimates.

Table A4.3 in the appendix shows that the relationship between inflation expectations and socioeconomic status is similar for the individuals in the mean-shift group with a few exceptions. For example, the coefficients on the *east*-dummy in Columns (2) and (4) are insignificant for the mean-shift group, while the coefficient in Column (8) is significant only at the 10% level. Similarly, education does not have a significant effect on histogram characteristics or point forecasts. Finally, the estimated effects of household income are larger and more significant in the regressions for the mean-shift group. These findings may hint at potential cross-sectional heterogeneity in the response of individuals when confronted with the alternative bin design. We analyze this issue in Section 4.3.3.

4.3.2 Differences in inflation expectations by treatment status

Having established the role of socioeconomic characteristics for inflation expectations, we now consider differences in expectations between the baseline group and the mean-shift group. Table 4.3 presents the estimates from linear regressions of inflation expectations on treatment status and socioeconomic characteristics (the latter are not shown) for the pooled sample of observations from both bin designs.

Table 4.3: Differences in inflation expectations across treatments.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.30***	-5.16***	-0.00	0.65***	-0.20***	0.05	-0.07	0.01	-0.59***
	(0.08)	(0.64)	(0.01)	(0.14)	(0.07)	(0.04)	(0.12)	(0.13)	(0.12)
Observations	2,733	2,733	2,733	2,733	2,733	1,473	1,473	2,678	2,678
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.05	0.02	0.02	0.02	0.01	0.01	0.04	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status and socioeconomic characteristics. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Differences in histogram characteristics

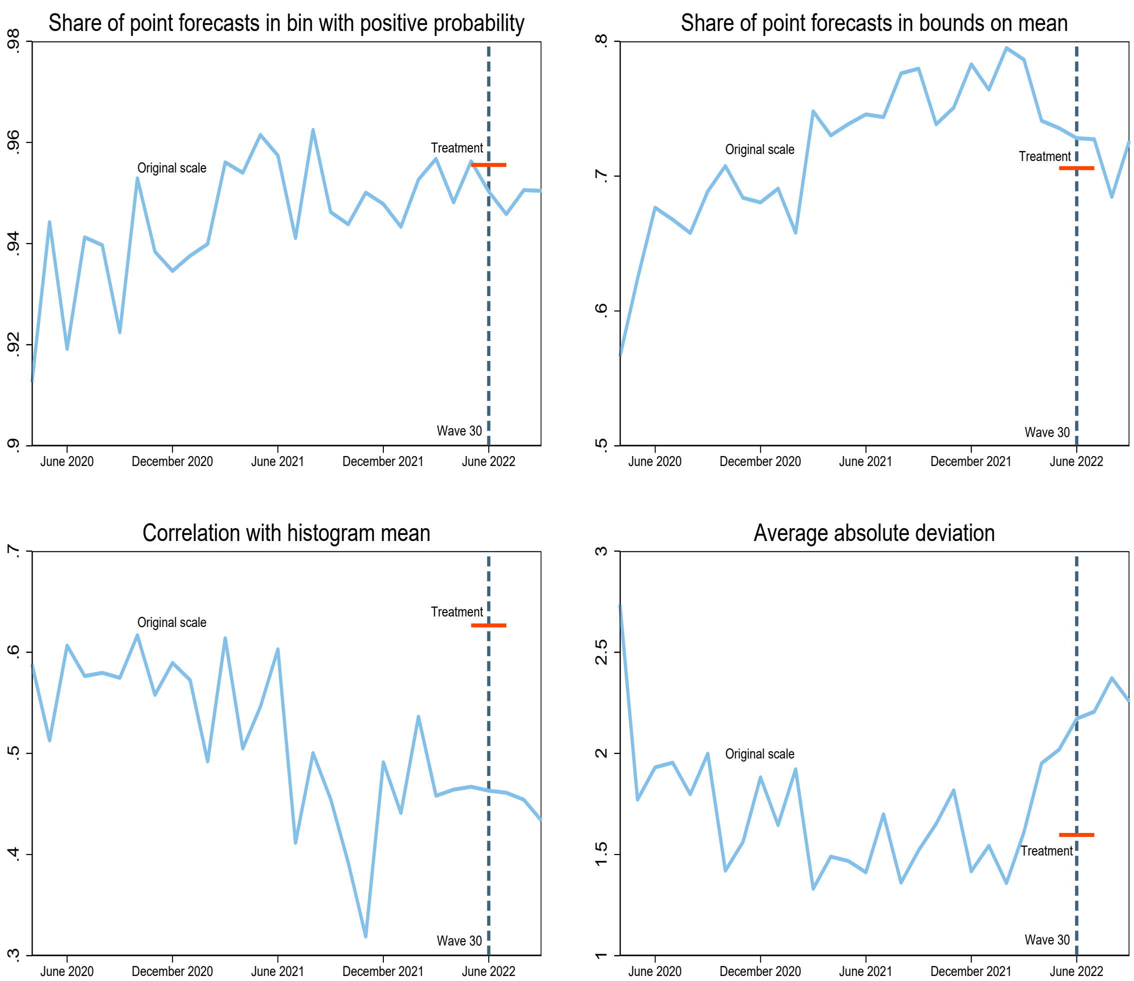
The histogram characteristics in Columns (1)-(7) are potentially affected by the alternative bin design. Indeed, we observe some noticeable differences in the histogram characteristics of both groups. In particular, we find that those in the mean-shift group use significantly more bins, assign a considerably lower probability mass to the rightmost bin, report higher histogram means and have lower inflation uncertainty than those in the baseline group. These effects are also economically significant. For example, Column (2) shows that the average probability mass in the rightmost bin is more than five percentage points lower for the mean-shift group than for the baseline group. This corresponds to the vertical difference in the right plot of Figure 4.1. Similarly, Column (4) shows that the histogram means in the mean-shift group are, on average, 0.65 percentage point higher than those in the baseline group.

Given that all other factors such as the macroeconomic environment or the remaining questions in the survey questionnaire were identical for all respondents, the observed differences in the histogram characteristics are either due to genuinely higher expectations, different discretization biases or framing effects. However, the upward shift in the average histogram mean remains below the upward shift in the bin definitions (0.65 percentage point versus four percentage points), suggesting that participants do not simply relocate their subjective distributions around the new center of the bin design. We provide a more detailed discussion of these issues in Section 4.4.

Consistency of point forecasts and probabilistic expectations

Bruine de Bruin et al. (2011) and Zhao (2022) find that the point forecasts of US households are

Figure 4.3: Consistency between point forecasts and probabilistic expectations.



Notes: For each BOP-HH wave, the upper-left plot shows the share of point forecasts that fall into a bin to which the respondent assigns nonzero probability. The upper-right plot depicts the share of point forecasts that lie within the bounds on the histogram mean. The lower-left plot presents correlations between point forecasts and histogram means. The lower-right plot shows the average absolute deviation between point forecasts and histogram means. Point forecasts are trimmed by 1% from top and bottom. The red bars are the corresponding figures for the mean-shift group in Wave 30.

well aligned with measures of central tendency such as the histogram mean. To assess whether this also is the case for German households, Figure 4.3 shows various measures of consistency between point forecasts and histogram means.

Around 95% of households report point forecasts that fall into a bin to which the respondents assigns nonzero probability. In line with the findings in Zhao (2022), approximately 70% of point forecasts lie within the individual bounds on the histogram mean, which are calculated by replacing the midpoint m_k in Equation (4.1) with the lower bound l_k and the upper bound u_k (see Engelberg et al., 2009, for details). These findings imply that point forecasts and probabilistic expectations of German households are relatively well-aligned and supplement the evidence for the US. However, the correlation between point forecasts and histogram means exhibits a declining trend over time. Similarly, the average absolute deviation between $\hat{\pi}_i^E$ and μ_i has increased in recent waves. These results suggest that the alignment between point forecasts and histograms suffers at times when households are forced to assign considerable probablity to the exterior bins as was the case in recent BOP-HH waves (see Figure 4.1). At the same time, the last two subfigures show a much higher degree of consistency for the mean-shift group in Wave 30. For example, the correlation between $\hat{\pi}_i^E$ and μ_i in Wave 30 is 0.63 for the mean-shift-group but only 0.46 for the baseline group. In light of these findings, we consider differences in the point forecasts and their alignment with the histogram means across treatment groups in the next step.

Column (8) of Table 4.3 shows that the point forecasts of individuals in the baseline group and the mean-shift groups are not significantly different from each other. In fact, the estimated coefficient on the *meanshift*-dummy is essentially zero. This is to be expected as the point forecast is elicited before the probabilistic expectation and respondents cannot return to this question later, this provides another confirmation that the randomization of treatments was successful. Our combined findings of significantly higher histogram means for the mean-shift group and stable point forecast across both groups suggest that the consistency between point forecasts and probabilistic expectations may be higher for one of the two groups. Indeed, Column (9) shows that the average absolute deviation between point forecasts and histogram means is significantly smaller in the mean-shift group. The effect size of almost 0.6 percentage point is economically relevant and similar in magnitude to the observed difference in the histogram means across groups.

In sum, the findings in columns (2), (4), (8) and (9) suggest that participants in the mean-shift group are able to more adequately communicate their higher probabilistic beliefs about future inflation. This, in turn, leads to a higher degree of consistency between the point forecasts and the probabilistic expectations reported by those individuals.

4.3.3 Heterogeneity in treatment effects

In this section, we analyze potential heterogeneity in the estimated treatment effects by including interaction terms between the treatment indicator and several characteristics of BOP-HH participants.

In a first step, we consider interactions of treatment status with socioeconomic characteristics. If households with different socioeconomic background react differently when presented with an alternative bin design, it may be recommendable to stick to the baseline design in order not to introduce additional distortions to the histogram characteristics. Tables A4.4-A4.10 in the appendix present the results. Overall, we find no evidence that the treatment effects significantly vary in the cross-section of households.

In a recent paper, Weber et al. (2022) notes that repeated participation may induce individuals to learn about a specific topic or details of the survey questionnaire. This effect is known as 'panel conditioning' and can also apply to the probabilistic expectations. Of the 2,733 households in our sample for Wave 30, 196 (7%) participated in the BOP-HH for the first time. 106 of these individuals are assigned to the baseline group and the other 90 to the mean-shift group. The remaining 2,537 individuals participated at least once before. The impact of our treatment on the probabilistic expectations may be stronger for more experienced survey participants. New entrants could simply assume that the alternative bin design represents the standard approach. On the other hand, it may be argued that participants with previous experience in the BOP-HH are somewhat 'anchored' around the original bin design. To explore these issues, we consider interactions between the *meanshift*-dummy and an indicator variable for first-time participants (*firsttimer*). Table 4.4 presents the results.

Table 4.4: Differences in inflation expectations: interaction with *firsttimer*.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.31***	-5.04***	0.01	0.68***	-0.17**	0.05	-0.12	-0.00	-0.63***
	(0.08)	(0.65)	(0.01)	(0.14)	(0.07)	(0.05)	(0.13)	(0.13)	(0.12)
firsttimer	0.78***	4.19*	0.11***	0.28	0.82***	-0.18*	-0.35	0.27	-0.07
	(0.23)	(2.17)	(0.04)	(0.33)	(0.22)	(0.09)	(0.22)	(0.35)	(0.29)
meanshift × *firsttimer*	-0.01	-0.96	-0.09*	-0.36	-0.37	-0.07	0.58	0.26	0.66
	(0.34)	(3.09)	(0.05)	(0.54)	(0.28)	(0.15)	(0.37)	(0.60)	(0.45)
Observations	2,733	2,733	2,733	2,733	2,733	1,473	1,473	2,678	2,678
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.05	0.03	0.02	0.02	0.02	0.01	0.04	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status, a dummy variable for first-time participants, an interaction with treatment status and socioeconomic characteristics. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

We find that, on average, first-time participants in Wave 30 assign nonzero probability to 0.78 more bins relative to more experienced respondents and assign over four percentage points of additional probability mass to the rightmost bin. As a result, new panelists provide more dispersed histograms that also tend to be more left-skewed than those of more households with more survey experience. They also have a higher probability of reporting multi-peaked probability distributions. In contrast, the point forecasts, histogram means and kurtosis of new entrants do not differ significantly from those of other participants. Importantly for our analysis, the interaction between *meanshift* and *firsttimer* is insignificant for all dependent variables except *multipeak*. This suggests that experienced participants do not react differently when presented with the alternative designs compared to new entrants who are confronted with questions about their probabilistic expectations for the first time.

In the last step, we consider interactions between treatment status and characteristics that capture the engagement of respondents with the survey. Table 4.5 shows the estimates of interacting treatment status with an indicator variable that states whether the respondent found the survey not interesting (*ninterest*), which is the case for 88 out of the 2,733 participants (3%)

Table 4.5: Differences in inflation expectations: interaction with *ninterest*.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	bins	*phigh*	*multipeak*	μ_i	σ_i	γ_i	κ_i	$\hat{\pi}_i^E$	$\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.27***	-4.94***	-0.00	0.72***	-0.23***	0.05	-0.05	0.01	-0.64***
	(0.08)	(0.63)	(0.01)	(0.14)	(0.07)	(0.04)	(0.12)	(0.13)	(0.12)
ninterest	-0.28	7.33	-0.05*	0.85	-0.55**	-0.23	0.80	-0.21	-0.78***
	(0.31)	(4.84)	(0.02)	(0.51)	(0.25)	(0.24)	(0.58)	(0.57)	(0.24)
meanshift × *ninterest*	0.94*	-6.47	0.04	-2.08**	0.75*	0.01	-0.59	-0.01	1.46**
	(0.55)	(5.45)	(0.05)	(0.87)	(0.40)	(0.30)	(0.73)	(0.95)	(0.71)
Observations	2,733	2,733	2,733	2,733	2,733	1,473	1,473	2,678	2,678
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.05	0.02	0.03	0.02	0.01	0.01	0.04	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status, a dummy variable that indicates whether respondents found the BOP-HH questionnaire uninteresting, an interaction with treatment status and socioeconomic characteristics. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

in Wave 30.

We find a significantly negative interaction between *meanshift* and *ninterest* in Column (4). This implies that while interested individuals in the mean-shift group report higher histogram means than the baseline group, the mean of uninterested individuals is, on average, more than one percentage point lower. In other words, those respondents tend to report lower inflation expectations than the baseline group despite the bins being moved towards higher inflation rates. As a result, the mismatch between point forecasts and histogram means tends to be *higher* for those individuals relative to the baseline group, whereas the opposite is the case for interested individuals in the treatment group. We also find that uninterested individuals in the baseline group express considerably lower inflation uncertainty than interested respondents in the baseline group. In contrast, uninterested individuals in the mean-shift group tend to report higher inflation uncertainty than individuals in the baseline group, whereas interested individuals in the mean-shift group tend to report lower standard deviations. In light of these findings, it may be recommendable to discard uninterested individuals from the sample altogether.[53]

We ran similar regressions with dummy variables that indicate whether the respondent found the survey too difficult (*difficult*, 8% of respondents) or too long (*toolong*, 22%). Tables A4.12-A4.13 present the estimates. While the results point in a similar direction as those for *ninterest*, the estimates are insignificant in most cases. However, we note that individuals that assign a high degree of difficulty to the survey questionnaire tend to report significantly different histogram moments than those who consider the survey as rather easy to answer. Moreover, for those individuals we also observe significant differences in the estimated treatment effects for higher moments such as skewness and kurtosis.

[53] Table A4.11 shows that our main results are very similar when focusing only on interested respondents.

4.3.4 Revisions of histogram moments

The rotating panel structure of the BOP-HH allows us to not only analyze differences in the point forecasts and probabilistic expectations in the cross-section of households, but also changes in revisions of such variables over time. In particular, we analyze i) how individuals who participated in Waves 29 to 31 updated their probabilistic expectations across time and ii) whether such revisions differ for those in the baseline group relative to those in the mean-shift group.

Of the 2,733 households in our sample for Wave 30, 738 also participated in Wave 29 and Wave 31. 368 of these respondents are in the baseline group and 370 in the mean-shift group. For those individuals we can compute revisions in point forecasts and histogram moments. For example, the revision of the histogram mean between Wave 29 and Wave 30 is defined as $\Delta\mu_i = \mu_{i,June} - \mu_{i,May}$. Similarly, $\Delta\mu_i = \mu_{i,July} - \mu_{i,June}$ is the corresponding revision between Wave 30 and Wave 31. The calculation for revisions of other variables proceeds analoguously. Table A4.14 in the appendix replicates Table 4.3 for the subset of respondents that participated in Waves 29 through 31. The estimates are very similar to our main results, although the magnitude of the effects tends to be slightly higher.

Updating from Wave 29 to Wave 30

While it is expected that some participants update their expectations from one period to the next, the magnitude of these changes can differ between treatment groups. In particular, if the differences between baseline and mean-shift groups in Table 4.3 can truly be ascribed to the treatment, the differences in revisions of histogram characteristics between Waves 29 and 30 should be similar in size to the estimated treatment effects. Table 4.6 presents the results from linear regressions of revisions of point forecasts and histogram moments on the treatment indicator variables and socioeconomic characteristics.

Table 4.6: Differences in revisions of inflation expectations between Wave 29 and Wave 30.

	(1) $\Delta bins$	(2) $\Delta phigh$	(3) $\Delta multipeak$	(4) $\Delta\mu_i$	(5) $\Delta\sigma_i$	(6) $\Delta\gamma_i$	(7) $\Delta\kappa_i$	(8) $\Delta\hat{\pi}_i^E$	(9) $\Delta\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.02	-3.94***	-0.03	0.63**	-0.42***	0.10	0.07	0.29	-0.54**
	(0.12)	(1.41)	(0.02)	(0.29)	(0.11)	(0.11)	(0.23)	(0.23)	(0.27)
Observations	738	738	738	738	738	386	386	713	713
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.00	0.03	0.01	0.00	0.03	-0.01	0.01	0.00	0.00

Notes: This table presents the estimates from linear regressions of revisions of histogram characteristics and point forecasts between Wave 29 and 30 on treatment status and socioeconomic characteristics for the subset of individuals that participated in Waves 29, 30 and 31. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

We find significant differences in revisions between the baseline group and the mean-shift group for the probability mass assigned to the rightmost bin, the histogram mean and the standard deviation. As expected, the differences in revisions between baseline and mean-shift group are closely associated with the size of the estimated treatment effects in Table 4.3. This further reinforces the notion that the observed differences can indeed be ascribed to the alternative bin design. Similarly, the coefficient on $\Delta|\hat{\pi}_i^E - \mu_i|$ is significant and the effect size is close to the corresponding estimate in Table 4.3.

Updating from Wave 30 to Wave 31

Next, we assess differences in revisions between Waves 30 and 31, i.e., immediately after we conducted our experiment. Individuals who were assigned to the mean-shift group in Wave 30 are now again presented with the baseline bin definitions. We are interested in the question of whether those individuals now revise their probabilistic expections as strongly in the opposite direction as they did when they were originally presented with the alternative bin design. Table 4.7 presents the estimates we obtain when replacing the revisions between Waves 29 and 30 with the revisions between 30 and 31. The socioeconomic characteristics are now drawn from Wave 31 instead of Wave 30.

Table 4.7: Differences in revisions of inflation expectations between Wave 30 and Wave 31.

	(1) $\Delta bins$	(2) $\Delta phigh$	(3) $\Delta multipeak$	(4) $\Delta\mu_i$	(5) $\Delta\sigma_i$	(6) $\Delta\gamma_i$	(7) $\Delta\kappa_i$	(8) $\Delta\hat{\pi}_i^E$	(9) $\Delta\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	-0.26**	7.11***	-0.01	-0.57**	0.25**	-0.09	0.10	-0.22	0.72***
	(0.10)	(1.39)	(0.02)	(0.28)	(0.10)	(0.12)	(0.31)	(0.22)	(0.26)
Observations	738	738	738	738	738	354	354	716	716
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.01	0.04	0.00	0.00	0.02	0.02	-0.01	0.01	0.01

Notes: This table presents the estimates from linear regressions of revisions of histogram characteristics and point forecasts between Wave 30 and 31 on treatment status and socioeconomic characteristics for the subset of individuals that participated in Waves 29, 30 and 31. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

When presented again with the original bin design, the individuals in the mean-shift group react by significantly reducing the number of bins, assigning considerably higher probability mass to the rightmost bin and reporting significantly lower histogram means and higher standard deviation. The significant estimates have the opposite sign as those in Table 4.6 and are similar in size. For the probability assigned to the rightmost bin and the misalignment between point forecasts and histogram means, the difference in revisions is even larger, which suggests that participants do not completely revert back to their pre-treatment expectations. Instead, they seem to partially retain their higher distribution from the mean-shift setting.

4.4 Discussion

We show that the mean-shift setting affects the probabilistic expectations of BOP-HH participants by allowing them to communicate more clearly their true beliefs at times when inflation is unusually high. However, other factors may also contribute to the observed deviations between treatment groups as discussed below.

One alternative explanation is that the differences in responses are driven by a central tendency bias, i.e., some respondents may believe that values close to the center of the distribution—zero for the baseline group, four for the mean-shift group—are deemed more likely by the Bundesbank (see Becker et al., 2023). Table 4.1 shows that the histograms in the baseline (mean-shift) group are centered around an average inflation rate of 6.57 (7.24) percent. These values are far away from the center of the respective distribution. Moreover, the difference in average histogram means of 0.67 percentage points is much smaller than the shift in the bins of four percentage points for the treatment group. These findings are more consistent with the interpretation that households are able to better state their true beliefs in the alternative setting rather than them using the center of the distribution as a focal point for their probabilistic expectations.

A second explanation is that at least some of the differences can be ascribed to different discretizations of the scale across treatments which affects histogram moments even under the assumption of stable beliefs. To explore the magnitude of such 'technical errors', we consider a hypothetical setting where a household with fixed probabilistic expectations is confronted with the two bin designs. The expectations of this household are normally distributed with known mean μ_0 and variance σ_0^2, i.e., $\mathcal{N}(\mu_0, \sigma_0^2)$. While it is unrealistic to assume that all households have normally distributed expectations, this may be an appropriate assumption for highly educated respondents. Also, Table 4.1 shows that the average skewness and kurtosis of BOP-HH participants are close to values expected under normality. For the mean, we choose $\mu_0 \in \{0, 4, 8\}$, where a value of zero corresponds to the center of the bin definitions for the baseline group, four corresponds to the center of the definitions for the mean-shift group and eight is close to the actual inflation rate in May 2022 (7.9%). For the variance, we consider $\sigma_0^2 \in \{4, 9\}$ to capture settings with low and high inflation uncertainty. For each combination of μ_0, σ_0^2 and the bin definitions, we calculate the probability mass assigned to each bin and compute the histogram moments using Equations (1)-(4). Table 4.8 presents the results. To faciliate the comparison between true and empirical moments, we report variances instead of standard deviations.

While the empirical histogram moments clearly deviate across settings, they are usually fairly close to the true values. The absolute difference between the empirical histogram means across treatments is at most 0.23 percentage point in case of the setting with low uncertainty and small values of μ_0. This is much smaller than the estimated difference of 0.65 percentage point between baseline and mean-shift group in Column (4) of Table 4.3. Turning to the variances, we observe that the empirical variances exceed their true value in all settings. The largest

Table 4.8: Histogram moments under stable expectations.

	Baseline group	Mean-shift group	Baseline group	Mean-shift group
	$\mathcal{N}(0,4)$		$\mathcal{N}(0,9)$	
μ	0.00	−0.23	0.00	−0.14
σ^2	4.77	4.85	10.46	9.87
γ	0.00	0.16	0.00	0.05
κ	3.61	2.80	3.03	3.07
	$\mathcal{N}(4,4)$		$\mathcal{N}(4,9)$	
μ	4.23	4.00	4.14	4.00
σ^2	4.85	4.77	9.87	10.46
γ	−0.16	0.00	−0.05	0.00
κ	2.80	3.61	3.07	3.03
	$\mathcal{N}(8,4)$		$\mathcal{N}(8,9)$	
μ	8.02	8.23	8.04	8.14
σ^2	5.24	4.85	9.52	9.87
γ	0.12	−0.16	0.02	−0.05
κ	2.24	2.80	2.74	3.07

Notes: For both bin definitions, this table presents the empirical histogram moments derived under the assumption that respondents have normally distributed inflation expectations.

difference between empirical variances across bin definitions—0.59 percentage point in absolute terms—is observed for the high-uncertainty scenario and small values of μ_0. This corresponds to an absolute difference in standard deviations of 0.09 percentage point. In contrast, Table 4.3 Column (5) shows that the estimated difference in standard deviations between treatment groups is more than twice as large. We conclude that our estimated treatment effects are too large to merely be the result of different discretizations across bin definitions.

4.5 Conclusion

For the current high-inflation environment, we find evidence that the moments of households' probabilistic inflation expectations vary with the response scale used to elicit them. In our sample, this is particularly the case for the histogram mean. As a result, the wedge between point forecast and histogram mean depends on the setup used for the probabilistic expectations. We show that the histogram variance is also affected. These findings do not appear to be the result of a central tendency bias or due to the use of different discretizations under the assumption of constant expectations. Rather, our results suggest that the inflation beliefs of German households have shifted upwards on average. Using the original scale to elicit expectations under these new beliefs tends to distort histogram moments as respondents have to allocate more probability mass to the higher, half-open interval in order to state their expectations. While we find the mean and variance to be affected, higher moments such as skewness and kurtosis appear to be relatively robust.

Our results have important implications for survey operators because they suggest that the interval design in household surveys could, and indeed should, be adjusted to the current macroeconomic environment as it is commonly done in surveys of professional forecasters. A more fine-grained interval design might also be advisable to accurately capture inflation expectations once inflation surges. However, such adjustments come at the cost of the comparibility across different household surveys. Another alternative would be to use sample splits such as the one used in this paper at times when inflation is unusually low or high. While some of the participants receive the original design to retain consistency with previous waves, the remaining panelists are confronted with an alternative design where the center of the bin is closer to the actual inflation rate.

Chapter 4 References

Armantier, O., Kosar, G., Pomerantz, R., Skandalis, D., Smith, K., Topa, G., & van der Klaauw, W. (2021). How economic crises affect inflation beliefs: Evidence from the Covid-19 pandemic. *Journal of Economic Behavior & Organization*, *189*, 443-469.

Armantier, O., Topa, G., der Klaauw, W. V., & Zafar, B. (2017). *An overview of the Survey of Consumer Expectations* (Staff Reports No. 800). Federal Reserve Bank of New York.

Becker, C., Duersch, P., & Eife, T. (2023). *Measuring inflation expectations: How stable are density forecasts?* (Mimeo)

Becker, C., Duersch, P., Eife, T., & Glas, A. (2022). *Extending the procedure of Engelberg et al. (2009) to surveys with varying interval-widths.* (FAU Discussion Papers in Economics No. 05/2022)

Beckmann, E., & Schmidt, T. (2020). *Bundesbank Online Pilot Survey on Consumer Expectations* (No. January 2020). Frankfurt (Main). (Technical Paper)

Bruine de Bruin, W., Manski, C. F., Topa, G., & van der Klaauw, W. (2011). Measuring consumer uncertainty about future inflation. *Journal of Applied Econometrics*, *26*(3), 454–478.

Bruine de Bruin, W., van der Klaauw, W., Downs, J., Fischhoff, B., Topa, G., & Armantier, O. (2010). Expectations of inflation: The role of demographic variables, expectation formation, and financial literacy. *Journal of Consumer Affairs*, *44*(2), 381-402.

Coibion, O., Gorodnichenko, Y., & Weber, M. (2022). Monetary policy communications and their effects on household inflation expectations. *Journal of Political Economy*, *130*(6), 1537-1584.

Conrad, C., Enders, Z., & Glas, A. (2022). The role of information and experience for households' inflation expectations. *European Economic Review*, *143*, 104015.

D'Acunto, F., Hoang, D., Paloviita, M., & Weber, M. (2021). *Human frictions in the transmission of economic policies* (NBER Working Paper No. 29279).

D'Acunto, F., Malmendier, U., Ospina, J., & Weber, M. (2021). Exposure to grocery prices and inflation expectations. *Journal of Political Economy*, *129*(5), 1615-1639.

Das, S., Kuhnen, C. M., & Nagel, S. (2020). Socioeconomic status and macroeconomic expectations. *The Review of Financial Studies*, *33*(1), 395-432.

Engelberg, J., Manski, C. F., & Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business & Economic Statistics*, *27*(1), 30–41.

Glas, A. (2020). Five dimensions of the uncertainty-disagreement linkage. *International Journal of Forecasting*, *36*(2), 607-627.

Glas, A., & Hartmann, M. (2022). Uncertainty measures from partially rounded probabilistic forecast surveys. *Quantitative Economics*, *13*(3), 979-1022.

Goldfayn-Frank, O., & Wohlfart, J. (2020). Expectation formation in a new environment: Evidence from the german reunification. *Journal of Monetary Economics*, *115*, 301-320.

Jonung, L. (1981). Perceived and expected rates of inflation in Sweden. *The American Economic Review*, *71*(5), 961–968.

Malmendier, U., & Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking? *The Quarterly Journal of Economics*, *126*(1), 373-416.

Malmendier, U., & Nagel, S. (2016). Learning from inflation experiences. *The Quarterly Journal of Economics*, *131*(1), 53-87.

Manski, C. F. (n.d.). Measuring expectations. *Econometrica*, *72*(5), 1329-1376.

Schwarz, N. (2010). Measurement as cooperative communication: What research participants learn from questionnaires. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The sage handbook of measurement* (pp. 43–61). Los Angeles: Sage.

Weber, M., D'Acunto, F., Gorodnichenko, Y., & Coibion, O. (2022). The subjective inflation expectations of households and firms: Measurement, determinants, and implications. *Journal of Economic Perspectives*, *36*(3), 157-84.

Zhao, Y. (2022). Internal consistency of household inflation expectations: Point forecasts vs. density forecasts. *International Journal of Forecasting*.

Chapter 4 Appendix

Table A4.1: Variable construction.

Variable	BOP-HH Questionnaire	Description
		Probabilistic inflation expectations
meanshift	*drandom2*	Equals one if the respondent belongs to the mean shift group (*drandom2*= 2), and zero for those in the baseline group (*drandom2*= 1).
bins	*infexprob_[a-j]* (CM004), *infexprob_rct1_[a-j]* (P3001A)	Number of bins to which the respondent assigns nonzero probability.
phigh	*infexprob_j* (CM004), *infexprob_rct1_j* (P3001A)	Probability mass assigned by the respondent to the highest available bin.
multipeak	same as *bins*	Equals one if the respondent provides a histogram with multiple peaks, and zero otherwise.
μ_i	same as *bins*	Mean of the histogram forecast for the German inflation rate over the next twelve months. We assume that the exterior bins have a width of four percentage points and that the probability mass in each bin is located at the midpoint.
σ_i	same as *bins*	Standard deviation of the histogram forecast.
γ_i	same as *bins*	Skewness of the histogram forecast.
κ_i	same as *bins*	Kurtosis of the histogram forecast.
		Point forecasts
$\hat{\pi}_i^P$	*devinfpoint* (CQ002)	Perceived German inflation rate over the previous twelve months in percent. This question was only asked to approximately one third of the participants in Wave 30.
$\hat{\pi}_i^E$	*infdef* (CM002) and *inflexppoint* (CM003)	Expected German inflation rate over the next twelve months in percent. Equals *inflexppoint* if *infdef* equals 'Inflation' and $(-1)\cdot$ *inflexppoint* if *infdef* equals 'Deflation'.
$\lvert\hat{\pi}_i^E - \mu_i\rvert$	same as $\hat{\pi}_i^E$ and μ_i	Absolute difference between the point forecast and the histogram mean.
		Socioeconomic characteristics
age	*age*	Age of individual. Set to 80 if *age* equals '80 years or older'.
east	*region*	Equals one if *region* equals 'east', and zero otherwise.
female	*gender*	Equals one if *gender* equals 'female', and zero otherwise.
fullemploy	*employ* (CS003)	Equals one if *employ* equals 'employed, full-time', and zero otherwise.
hhsize	*hhsize* (CS006)	Household size. Set to 6 if *hhsize* equals '6 or more'.
income	*hhinc* (CS008)	Monthly household income in €1,000 (using bin midpoints): = 0.25 if *hhinc* equals 'Less than €500', = 0.75 if *hhinc* equals '€500 to €999', = 1.25 if *hhinc* equals '€1,000 to €1,499', = 1.75 if *hhinc* equals '€1,500 to €1,999', = 2.25 if *hhinc* equals '€2,000 to €2,499', = 2.75 if *hhinc* equals '€2,500 to €2,999', = 3.25 if *hhinc* equals '€3,000 to €3,499', = 3.75 if *hhinc* equals '€3,500 to €3,999', = 4.50 if *hhinc* equals '€4,000 to €4,999', = 5.50 if *hhinc* equals '€5,000 to €5,999', = 7.00 if *hhinc* equals '€6,000 to €7,999', = 9.00 if *hhinc* equals '€8,000 to €9,999', = 11.00 if *hhinc* equals '€10,000 or more'.

Notes: This table describes the construction of the variables used in the empirical analysis. In the middle column, we refer to the names of the original variables as listed in the questionnaire for Wave 30 (June 2022) of the BOP-HH.

Table A4.1: Variable construction (continued).

Variable	BOP-HH Questionnaire	Description
yoe	*eduschool* (CS001)	Years of education of individual following SOEP definition: $\begin{cases} = 7 \text{ if } \textit{eduschool} \text{ equals 'No school-leaving certificate'}, \\ = 9 \text{ if } \textit{eduschool} \text{ equals 'Secondary school-leaving certificate'}, \\ = 10 \text{ if } \textit{eduschool} \text{ equals 'Other school-leaving certificate'}, \\ = 10 \text{ if } \textit{eduschool} \text{ equals 'Intermediate secondary school certificate'}, \\ = 10 \text{ if } \textit{eduschool} \text{ equals 'Polytechnical secondary school certificate (8th/10th grade)'}, \\ = 13 \text{ if } \textit{eduschool} \text{ equals 'University of applied sciences entrance diploma / completed technical school'}, \\ = 13 \text{ if } \textit{eduschool} \text{ equals 'Senior school-leaving certificate/ general or subject-specific university entrance diploma'}, \\ = 18 \text{ if } \textit{eduschool} \text{ equals 'College / university degree'}. \end{cases}$
		Aditional characteristics
firsttimer	*id*	Equals one if the respondent participated in the BOP-HH for the first time in Wave 30, and zero otherwise.
ninterest	*qinterest*	Equals one if the respondent found the BOP-HH ‘not so interesting’ or ‘not interesting at all’, and zero otherwise.
difficult	*qeasy*	Equals one if the respondent found the BOP-HH ‘somewhat difficult’ or ‘very difficult’, and zero otherwise.
toolong	*qlong*	Equals one if the respondent found the BOP-HH ‘a little too long’ or ‘far too long’, and zero otherwise.

Notes: This table describes the construction of the variables used in the empirical analysis. In the middle column, we refer to the names of the original variables as listed in the questionnaire for Wave 30 (June 2022) of the BOP-HH.

Table A4.2: Treatment assignment and socioeconomic characteristics.

	meanshift
age	-0.06
	(0.08)
east	-0.74
	(2.57)
female	2.64
	(2.03)
fullemploy	-2.02
	(2.36)
hhsize	-0.17
	(1.09)
ln(*income*)	0.51
	(2.12)
yoe	-0.35
	(0.60)
Constant	53.91***
	(16.52)
Observations	2,733
$\bar{R}^2$	0.00

Notes: This table presents the estimates from a linear regression of treatment status on socioeconomic characteristics. The baseline group consists of the individuals that were presented with the original bin design. Heteroskedasticity-consistent standard errors are reported in parentheses. The reported coefficients and standard errors are the estimated ones times 100. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Table A4.3: Inflation expectations and socioeconomic characteristics: mean-shift group.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
age	-0.02***	0.00	0.00***	0.01	-0.01**	-0.00	0.02***	-0.00	-0.00
	(0.01)	(0.04)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
east	-0.10	0.87	0.02	0.25	0.02	0.08	0.06	0.53*	0.11
	(0.17)	(1.07)	(0.02)	(0.25)	(0.14)	(0.08)	(0.19)	(0.28)	(0.20)
female	-0.16	2.19***	0.03**	0.76***	0.16	0.01	-0.15	0.99***	0.43**
	(0.13)	(0.77)	(0.02)	(0.20)	(0.11)	(0.06)	(0.17)	(0.21)	(0.17)
fullemploy	0.13	2.53**	0.01	0.32	0.09	-0.06	-0.01	0.17	0.04
	(0.15)	(1.01)	(0.01)	(0.24)	(0.11)	(0.06)	(0.15)	(0.23)	(0.19)
hhsize	-0.00	0.92**	0.02**	0.19	0.03	-0.07**	0.09	0.35***	0.25**
	(0.07)	(0.43)	(0.01)	(0.12)	(0.05)	(0.03)	(0.07)	(0.11)	(0.12)
ln(*income*)	-0.29*	-2.64***	-0.05***	-0.76***	-0.25*	0.07	-0.50**	-0.93***	-0.46*
	(0.16)	(0.96)	(0.02)	(0.21)	(0.13)	(0.07)	(0.21)	(0.25)	(0.26)
yoe	0.07**	0.03	-0.01**	-0.06	0.01	0.02	0.08	-0.05	-0.02
	(0.04)	(0.23)	(0.00)	(0.06)	(0.03)	(0.02)	(0.05)	(0.06)	(0.05)
Constant	6.18***	19.68***	0.39***	12.87***	4.09***	-0.46	5.32***	15.14***	4.80**
	(1.24)	(6.47)	(0.14)	(1.53)	(1.01)	(0.55)	(1.54)	(2.20)	(2.33)
Observations	1,377	1,377	1,377	1,377	1,377	778	778	1,350	1,350
$\bar{R}^2$	0.03	0.02	0.02	0.02	0.01	0.00	0.02	0.05	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on socioeconomic characteristics. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Table A4.4: Differences in inflation expectations: interaction with *age*.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.44	-7.60***	-0.02	1.28**	-0.34	0.04	-0.16	0.65	-0.80
	(0.33)	(2.59)	(0.04)	(0.57)	(0.25)	(0.15)	(0.42)	(0.55)	(0.51)
meanshift × *age*	-0.00	0.04	0.00	-0.01	0.00	0.00	0.00	-0.01	0.00
	(0.01)	(0.04)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
Observations	2,733	2,733	2,733	2,733	2,733	1,473	1,473	2,678	2,678
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.05	0.02	0.03	0.01	0.01	0.01	0.04	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status, socioeconomic characteristics and an interaction between treatment status and age. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Table A4.5: Differences in inflation expectations: interaction with *east*.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.29***	-4.63***	-0.01	0.73***	-0.23***	0.04	-0.07	0.04	-0.63***
	(0.09)	(0.65)	(0.01)	(0.15)	(0.07)	(0.05)	(0.13)	(0.14)	(0.13)
meanshift × *east*	0.05	-3.06	0.04	-0.45	0.17	0.05	-0.01	-0.15	0.21
	(0.22)	(2.04)	(0.03)	(0.36)	(0.19)	(0.12)	(0.39)	(0.38)	(0.29)
Observations	2,733	2,733	2,733	2,733	2,733	1,473	1,473	2,678	2,678
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.05	0.02	0.03	0.02	0.01	0.01	0.04	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status, socioeconomic characteristics and an interaction between treatment status and a dummy variable for East Germans. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Table A4.6: Differences in inflation expectations: interaction with *female*.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.31***	-3.93***	0.01	0.54***	-0.20***	-0.02	-0.04	-0.03	-0.53***
	(0.10)	(0.68)	(0.01)	(0.16)	(0.07)	(0.05)	(0.15)	(0.15)	(0.13)
meanshift × *female*	-0.03	-3.25**	-0.03	0.31	-0.01	0.20**	-0.10	0.11	-0.17
	(0.17)	(1.43)	(0.02)	(0.30)	(0.15)	(0.09)	(0.25)	(0.28)	(0.25)
Observations	2,733	2,733	2,733	2,733	2,733	1,473	1,473	2,678	2,678
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.05	0.02	0.02	0.01	0.01	0.01	0.04	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status, socioeconomic characteristics and an interaction between treatment status and a dummy variable for female respondents. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Table A4.7: Differences in inflation expectations: interaction with *fullemploy*.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.18*	-5.71***	-0.00	0.62***	-0.28***	0.10	0.04	0.06	-0.51***
	(0.11)	(0.80)	(0.01)	(0.18)	(0.10)	(0.06)	(0.19)	(0.15)	(0.15)
meanshift × *fullemploy*	0.27*	1.29	0.01	0.07	0.18	-0.11	-0.24	-0.12	-0.19
	(0.16)	(1.31)	(0.02)	(0.28)	(0.13)	(0.09)	(0.24)	(0.27)	(0.23)
Observations	2,733	2,733	2,733	2,733	2,733	1,473	1,473	2,678	2,678
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.05	0.02	0.02	0.02	0.01	0.01	0.04	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status, socioeconomic characteristics and an interaction between treatment status and a dummy variable for full-time employed individuals. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Table A4.8: Differences in inflation expectations: interaction with *hhsize*.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.47**	-4.84***	-0.03	0.64*	-0.17	0.05	-0.03	-0.43	-0.91***
	(0.19)	(1.59)	(0.02)	(0.34)	(0.16)	(0.10)	(0.29)	(0.32)	(0.29)
meanshift × *hhsize*	-0.08	-0.15	0.01	0.01	-0.02	-0.00	-0.02	0.20	0.15
	(0.08)	(0.70)	(0.01)	(0.15)	(0.06)	(0.04)	(0.10)	(0.14)	(0.13)
Observations	2,733	2,733	2,733	2,733	2,733	1,473	1,473	2,678	2,678
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.05	0.02	0.02	0.01	0.01	0.01	0.04	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status, socioeconomic characteristics and an interaction between treatment status and household size. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Table A4.9: Differences in inflation expectations: interaction with ln(*income*).

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	2.25	-14.91	0.08	2.72	1.00	0.64	4.11**	-1.08	-1.54
	(1.38)	(11.08)	(0.15)	(2.25)	(1.16)	(0.72)	(2.09)	(2.32)	(2.24)
meanshift × ln(*income*)	-0.24	1.20	-0.01	-0.25	-0.15	-0.07	-0.51**	0.13	0.12
	(0.17)	(1.34)	(0.02)	(0.27)	(0.14)	(0.09)	(0.25)	(0.28)	(0.27)
Observations	2,733	2,733	2,733	2,733	2,733	1,473	1,473	2,678	2,678
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.05	0.02	0.02	0.02	0.01	0.02	0.04	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status, socioeconomic characteristics and an interaction between treatment status and household income. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Table A4.10: Differences in inflation expectations: interaction with *yoe*.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	-0.13	-14.20***	-0.00	0.86	-0.44	0.52*	0.63	-1.24	-1.49*
	(0.59)	(4.91)	(0.08)	(0.99)	(0.52)	(0.31)	(0.88)	(0.95)	(0.84)
meanshift × *yoe*	0.04	0.78*	0.00	-0.02	0.02	-0.04	-0.06	0.11	0.08
	(0.05)	(0.42)	(0.01)	(0.08)	(0.04)	(0.03)	(0.07)	(0.08)	(0.07)
Observations	2,733	2,733	2,733	2,733	2,733	1,473	1,473	2,678	2,678
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.05	0.02	0.02	0.01	0.01	0.01	0.04	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status, socioeconomic characteristics and an interaction between treatment status and years of education. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Table A4.11: Differences in inflation expectations: interested participants only.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.27***	-4.94***	-0.00	0.72***	-0.23***	0.05	-0.05	0.01	-0.63***
	(0.08)	(0.63)	(0.01)	(0.14)	(0.07)	(0.04)	(0.12)	(0.13)	(0.12)
Observations	2,645	2,645	2,645	2,645	2,645	1,425	1,425	2,592	2,592
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.03	0.05	0.02	0.03	0.02	0.01	0.01	0.04	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status and socioeconomic characteristics when focusing only on participants who find the BOP-HH interesting. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Table A4.12: Differences in inflation expectations: interaction with *difficult*.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.30***	-5.00***	-0.00	0.70***	-0.21***	0.08*	-0.13	0.00	-0.65***
	(0.08)	(0.67)	(0.01)	(0.14)	(0.07)	(0.05)	(0.13)	(0.13)	(0.12)
difficult	0.61***	-0.37	0.04	-0.67*	0.50**	0.14*	-0.37*	-0.61*	-0.18
	(0.24)	(1.83)	(0.03)	(0.36)	(0.20)	(0.09)	(0.20)	(0.33)	(0.28)
meanshift × *difficult*	0.06	-2.25	0.04	-0.77	0.13	-0.32**	0.66**	0.05	0.82*
	(0.35)	(1.94)	(0.05)	(0.50)	(0.30)	(0.14)	(0.33)	(0.52)	(0.49)
Observations	2,733	2,733	2,733	2,733	2,733	1,473	1,473	2,678	2,678
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.05	0.03	0.03	0.02	0.01	0.01	0.05	0.02

S

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status, a dummy variable that indicates whether respondents found the BOP-HH questionnaire too difficult, an interaction with treatment status and socioeconomic characteristics. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Table A4.13: Differences in inflation expectations: interaction with *toolong*.

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.32***	-5.06***	-0.00	0.79***	-0.22***	0.05	-0.11	0.03	-0.66***
	(0.09)	(0.70)	(0.01)	(0.15)	(0.07)	(0.05)	(0.13)	(0.15)	(0.13)
toolong	0.22*	1.59	0.01	0.20	0.10	-0.07	0.16	-0.03	-0.16
	(0.13)	(1.31)	(0.02)	(0.24)	(0.11)	(0.08)	(0.22)	(0.22)	(0.21)
meanshift × *toolong*	-0.05	-0.31	-0.00	-0.63*	0.08	-0.03	0.22	-0.11	0.35
	(0.20)	(1.61)	(0.02)	(0.34)	(0.17)	(0.12)	(0.32)	(0.32)	(0.28)
Observations	2,733	2,733	2,733	2,733	2,733	1,473	1,473	2,678	2,678
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.05	0.02	0.03	0.02	0.01	0.01	0.04	0.02

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status, a dummy variable that indicates whether respondents found the BOP-HH questionnaire too long, an interaction with treatment status and socioeconomic characteristics. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Table A4.14: Differences in inflation expectations: Wave 29 to Wave 31 participants

	(1) *bins*	(2) *phigh*	(3) *multipeak*	(4) μ_i	(5) σ_i	(6) γ_i	(7) κ_i	(8) $\hat{\pi}_i^E$	(9) $\lvert\hat{\pi}_i^E - \mu_i\rvert$
meanshift	0.09	-5.22***	-0.01	0.83***	-0.37***	0.09	0.21	-0.13	-0.97***
	(0.14)	(1.08)	(0.02)	(0.28)	(0.12)	(0.08)	(0.21)	(0.25)	(0.24)
Observations	738	738	738	738	738	405	405	722	722
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\bar{R}^2$	0.04	0.06	0.01	0.04	0.04	0.03	0.02	0.06	0.03

Notes: This table presents the estimates from linear regressions of histogram characteristics and point forecasts on treatment status and socioeconomic characteristics for the subset of individuals that participated in Waves 29, 30 and 31. The baseline group consists of the individuals that were presented with the original bin design. In columns (6) and (7), we consider only the responses of participants who use at least three bins. For columns (8) and (9), we trim $\hat{\pi}_i^E$ by 1% from top and bottom. Heteroskedasticity-consistent standard errors are reported in parentheses. Asterisks '*', '**', and '***' indicate significance at the 10%, 5%, and 1% critical level, respectively.

Chapter 5

Marriage, Parenthood and Social Network: Subjective Well-Being and Mental Health in Old Age

Abstract[‖]

Parenthood, marital status and social networks have been shown to relate to the well-being and mental health of older people. Using a large sample of respondents aged 50 and older from 16 European countries, we identify the associations of well-being and mental health with family status. Making use of detailed social network data of the respondents, we also identify how different social support networks correlate with the well-being and health indicators. We observe positive associations for all network types, over and beyond any direct associations of family status with well-being. Results suggest that non-residential children are important providers of social support for their parents at older age.

[‖] Joint work with Stefan T. Trautmann and Isadora Kirchmaier. The content of this chapter has been published as: Becker, C., Kirchmaier, I., & Trautmann, S. T. (2019). Marriage, parenthood and social network: Subjective well-being and mental health in old age. PloS one, 14(7), e0218704.

This paper uses data from SHARE Waves 1, 2, and 4, (DOIs: 10.6103/SHARE.w1.600, 10.6103/SHARE.w2.600, 10.6103/SHARE.w4.600), see Börsch-Supan et al. (2013) for methodological details. For funding of SHARE, see the Funding section.

We thank Luisa Kling for her excellent research assistance, as well as Christian König-Kersting and Martin Vollmann for comments that greatly improved the manuscript.

5.1 Introduction

The link between family status (marital status and parenthood), well-being, and mental health is widely discussed in academic and popular discourses. Evidence suggests that being married or living with a partner can have a positive effect on life satisfaction (Mastekaasa, 1994) and is associated with higher well-being, better mental health and fewer depressive symptoms in old age (Buber & Engelhardt, 2008; Bures et al., 2009; Gibney et al., 2017).

Parenthood, on the other hand, does not appear to be associated with enhanced mental health (Evenson & Simon, 2005; Hansen et al., 2009; Hansen, 2012). The risk of depression is especially pronounced for women with parenting stress and poor physical health, but less pronounced for those being supported by the partner (Manuel et al., 2012). Repeated cross-sectional data on US parents and non-parents shows a gap in subjective well-being between these two groups, which, however, becomes smaller over the period 1973 through 2008 due to decreased happiness of non-parents (Herbst & Ifcher, 2016). A cross-country comparison finds only weak associations between life satisfaction and having children, with unclear direction (Mastekaasa, 1994). However, there is also evidence that the relationship between children and well-being becomes more positive for older respondents (Mastekaasa, 1994; Margolis & Myrskylä, 2011). Depending on the life-cycle stage, the aspects of parenthood may thus differ, suggesting that the positive aspects of parenthood dominate when getting older. Amongst others, the role of children as a form of social support may become important in the later stages of a person's life (Margolis & Myrskylä, 2011).

But what constitutes social support? One of the most cited definitions stems from Cobb (1976), describing social support as "information leading the subject to believe that he is cared for and loved, esteemed and a member of mutual obligations". The US National Cancer Institute (*Definition of social support - NCI Dictionary of Cancer Terms - National Cancer Institute*, 2023), defines social support as "a network of family, friends, neighbors, and community members that is available in times of need to give psychological, physical, and financial help". In general, a social network consists of a "set of actors and the ties amongst them" (Wasserman & Faust, 1994), while the term social support further describes the quantity and quality of these ties from an individual perspective. While there exist multiple definitions of social support, most of them encompass factors for the size and structure of the network, as well as including measures for physical distance to other network members, length of the relationships, frequency of contact or function of each relationship (Pearson, 1986). Evidence suggests that such social support networks are related to less loneliness and more happiness (Litwin & Shiovitz-Ezra, 2011; Litwin & Stoeckel, 2014) and act as important buffer against stressful events (Cobb, 1976; Pearson, 1986).

While results on parenthood might be controversial and depend on the age of the studied population, there is widespread agreement that social support is associated with higher life satisfaction, and that social networks are an important factor for well-being (Pinquart & Sörensen, 2000). Bringing these two branches of the literature together, we aim to shed light on the link between a person's family status, the resulting characteristics of their social networks, and their well-being and mental health, using a large sample of 55.000 middle-aged and older adults from 16 European countries. This sample was taken from the Survey of Health, Aging and Retirement in Europe (SHARE). For people aged 60 and older in Mediterranean and non-Mediterranean countries in the first wave of the data set, there is some evidence that the number of residential children is associated with more depressive symptoms (Litwin, 2010). We aim to expand and generalize these findings using recently collected, detailed network data, across European countries. Parenthood, marital status and different types of social networks might help to sustain well-being and mental health in old age. Our objective in the current study is to analyze the role of marital status, parenthood and social networks deriving from these family backgrounds as potential sources of social support, for well-being and mental health in old age. We consider four distinct measures, used in different fields such as economics or psychology, to obtain a comprehensive picture of well-being and mental health. These are the CASP-12 scale for quality of life, the EURO-D scale for depressive symptoms, and one question each for life and social support network satisfaction.

We use the full range of the SHARE data set, which includes people aged 50 and older. At this point in the life cycle, parents may have resident children, children living away from home, and grandchildren, allowing us to separate their associations with well-being. We use network composition measures in order to determine network types, and control for network size and relational dynamics separately. Additionally, we calculate the network types for each country separately, taking cultural differences in network compositions into account.

Based on the current literature we test the following three hypotheses for the well-being and mental health of people aged 50 and older: i) A positive association with being married, ii) a positive association with the number of children and grandchildren not living at home, and iii) a positive association with having a strong social network implied by family background. We proceed as follows: Section 5.2 describes the data used and our methods to measure well-being, mental health and the characteristics of social support networks in detail. In Section 5.3 we present the results of our analysis. We first analyze the association of family status with well-being and mental health measures without taking the social network into account. We then take the network composition as criterion variables and use hierarchical clustering to determine social network types which differ mainly in their main source of social support. We then assess the relationship between the resulting social support network types and outcome measures, controlling for family status, network size, and relational dynamics. Section 5.4 discusses our findings and provides concluding remarks.

5.2 Data and Methods

5.2.1 Respondents

We use data from the cross-national panel database Survey of Health, Aging and Retirement in Europe (SHARE), release 6.0.0., managed by the Munich Center for the Economics of Aging, Max Planck Institute for Social Law and Social Policy (Börsch-Supan et al., 2013, 2015; Malter & Börsch-Supan, 2013). The cross-national panel database provides extensive data on health and socio-economic status. The target population is people of age 50 or older having their regular domicile in the respective country. Current partners are interviewed regardless of their age. We make use of SHARE wave 4 (Börsch-Supan, 2018c) that was administered between 2010 and 2012 in 16 European countries, and includes a module on social network. We update missing constants with data from waves 1 and 2 (Börsch-Supan, 2018a,b). We include respondents age 50 and older not living in a nursing home. The number of respondents differs by country. Over all countries, there are about 55.000 observations available. For an overview of the total number of observations of each country, see the supporting Table A5.1.

5.2.2 Demographic factors

The SHARE data set contains detailed data on demographics. Summary statistics of all demographic variables used in the analyses can be found in Table A5.2. The demographic factor of interest is the family status, which we measure by the marital status, total number of children, children living at home, and grandchildren. Over all countries 70% of the respondents are married and 91% have children.

The marital status of each respondent is classified into the categories (1) married and living together with spouse, (2) registered partnership, (3) married and living separated from spouse, (4) never married, (5) divorced, (6) widowed. For the regression analysis we construct the dummy variables *married* which takes the value of one if the respondent is married or in a registered partnership, the dummy variable *divorced*, which takes the value of one if the respondent is divorced or living separated from spouse, and a dummy variable *widowed*. We include respondents living separately from their spouse in the dummy *divorced*, as living separately is often a preceding step to a divorce.

Parenthood is measured by the number of children alive and the number of resident children, including fostered, adopted and stepchildren. We define the four-category measure *children* with categories no children, one child, two children, and three or more children, and create the respective dummy variables for each category. We further construct the variables *resident children* and *grandchildren* which report for each respondent the number of children living with the family and the number of grandchildren.

Further demographics are used as controls. The set *Controls A* consist of gender, age (of the respondent at the time of the interview), age squared, and a dummy variable indicating the country of residence of the respondent to control for cultural differences. The set *Controls B* additionally includes dummies for urban character of residence, being employed, self-employment, level of education according to the international classification of education ISCED-97 ("International Standard Classification of Education, ISCED 1997", 2003), an indicator for the average monthly household income, and the aforementioned dummies for divorced and widowed. In SHARE wave 4, each household respondent is asked to state the overall after-tax income of the entire household in an average month of last year. If a respondent refuses to answer, the interviewer asks whether the respondent earns more, less or approximately the amount in certain bracketed values, which represent country-specific 25th, 50th, and 75th percentiles of the reported household incomes from SHARE wave 2. We use the information from the stated household income and the unfolding brackets and define four categories for the average monthly household income: (1) Low income [0 to 25th percentile], (2) Middle income [25th to 50th percentile], (3) Upper middle income [50th percentile to 75th percentile], and (4) High income [75th percentile and higher]. The boundaries of the intervals are the country-specific bracket values of SHARE wave 4 (details and summary statistics in Table A5.2).

In order to control for health, we include a measure of self-assessed physical health (Would you say your health is: (1) poor, (2) fair, (3) good, (4) very good, and (5) excellent), and whether drugs for sleeping problems, anxiety or depression are taken.

5.2.3 Well-Being and Mental Health Indicators

Well-being can be defined as the psychological balance point between individually available resources and challenges (Dodge et al., 2012) and may be linked to many different aspects of life. In order to develop national well-being measures, the Office for National Statistics in the UK ran a public debate on the question through various platforms (Matheson, 2011). The three most frequent answers to the question "What things matter most in your life? What is Well-being?" were "Health", "Having good connections with friends and relatives", and "Job satisfaction (and economic security)" (Evans, 2011). Many empirical studies report a link between socioeconomic status, quality and quantity of social contacts, and well-being (Pinquart & Sörensen, 2000). In our study, we use a broad set of measures to map respondents' well-being: a simple single-item question regarding life satisfaction; the CASP-12 multi-item quality of life scale; a single-item question on social support network satisfaction; and the EURO-D depressive symptoms scale. In the following, we will discuss the three measures in more detail. We also use measures of health, education, and financial status as controls in our analyses (Diener & Suh, 1997; Knesebeck et al., 2007).

The first measure concerns a general feeling about the quality of life, the stated *Life satisfaction*. It is extracted by a single-item question in which respondents indicate on a scale from 0 (low satisfaction) to 10 (high satisfaction) how satisfied they are with their life. This scale has acceptable reliability and validity (Pavot & Diener, 1993; Beckie & Hayduk, 1997).

The second measure is the CASP-12, *quality of life scale*, which is designed to capture quality of life in old age (Hyde et al., 2003). Participants indicate for twelve statements whether they apply on a scale from 1 (often) to 4 (never). The twelve questions concern four dimensions of quality of life, control, autonomy, pleasure and self-realization, resulting in an aggregate index ranging from 12 (low quality of life) to 48 (high quality of life). We normalize it such that it ranges from 0 (low quality of life) to 10 (high quality of life).

The third measure concerns the stated *Network satisfaction*. Respondents indicate on a scale from 0 (low satisfaction) to 10 (high satisfaction) how satisfied they are with their social network. If respondents indicated that there is no person with whom they discuss matters or there is no one who is important to them, they were asked how satisfied they were with this fact.

The fourth measure is the EURO-D depression score (Prince et al., 1999). It is an indicator for depressive symptoms and captures aspects of mental health in late life. It has been demonstrated to provide a valid comparison of depressive symptoms across European countries (Prince et al., 1999; Castro-Costa et al., 2008). The EURO-D depression score is generated from questions on 12 dimensions: Depression, pessimism, suicidality, guilt, sleep, interest, irritability, appetite, fatigue, concentration, enjoyment, and tearfulness. Respondents are asked whether there is an indication for each of these dimensions. It results in an aggregate index ranging from 0 (not depressed) to 12 (very depressed). We normalize it such that it ranges from 0 (very depressed) to 10 (not depressed) and call it *Lack of depressive symptoms*.

Figure 5.1 presents the average of the well-being measures at each age until 91 (see Figure A5.1 for the age distribution). While network satisfaction and life satisfaction remain relatively stable, the quality of life index and lack of depressive symptoms index decline beyond age 65. The graphs for male and female respondents are rather similar, except for the lack of depressive symptoms index; male respondents have on average a 0.73 points higher index ($p < 0.01$, Mann-Whitney-U test; Figure A5.2).

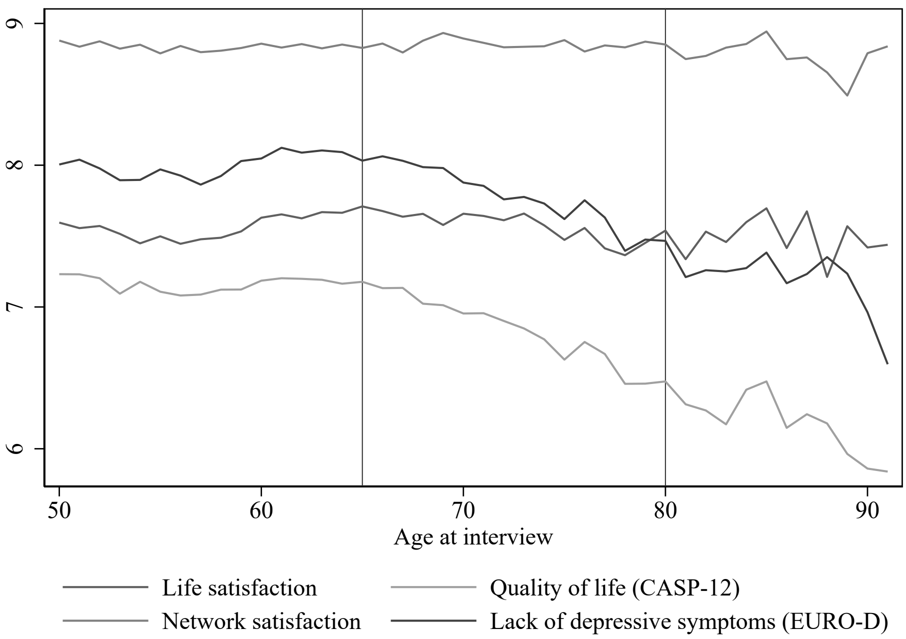
5.2.4 Social Support Networks

A social support network can be characterized by its size and composition (percentage of partner, children, other relatives, and friends in the network) and relational dynamics. In Wave 4, the SHARE respondents are asked to answer questions about their social support network along the dimensions (1) size, (2) relationship, (3) contact frequency, (4) proximity, and (5) closeness.

In order to identify the members of their *social support network*, the respondents were asked to mention the name of persons with whom they discuss important matters. The total number of persons in the social support network is its size. It is possible to mention up to seven persons, however this boundary is only mentioned if it is reached. Only 3% of the respondents reach this boundary. Most respondents state one, two or three persons as members of their social support network (28%, 25%, and 20% of the respondents, respectively). Evidence suggests that the number of network members is positively linked with life satisfaction (Tomini et al., 2016), but that in old age the network is reduced to members with close contact (Fung et al., 2001).

The *composition* of a network refers to the relationship type between each member. A

Figure 5.1: Average well-being and mental health measure.



Notes: Average well-being and mental health measure for all ages from 50 to 90 years. After age 91 the number of available observations drops to less than 50.

person who has daily contact with two children and a person who has daily contact with two friends have a social support network of equal size and contact frequency, however, they have a different main source of social support. In a meta-analysis, Pinquart and Sörensen (2000) provide evidence that the quantity of social contacts with friends is more strongly related with subjective well-being than the quantity of social contacts with family. They argue that friends are voluntary relationships, and they are typically members of the same age group or share similar preferences. Still, especially in older age, spouses and children are a crucial part of networks. Later in life, parents desire open communication, but low interference in each other's lives thereby maintaining independence in old age and minimizing intergenerational conflicts (Blieszner & Mancini, 1987). Brandt et al. (2009) analyzed the type of support between older parents, their children and professional providers. They found that children play a central role in providing help for their parents in the household and with paperwork. In Southern Europe, they are more likely to also take over regular medical care. There can also be differences within family structures. Shanas (1979) provides evidence that the immediate family (partner and children) is the major social support during illness, and the extended family (children, siblings, and other relatives) is the tie to the community.

There are different ways to determine these different types of networks. One way is to construct network types, in which people are similar along family status (e.g. marital status, number of children and close relatives) and network measures (e.g. number of close friends, frequency of contact with family and friends, and frequency of attending social events). Commonly, there are four to five network types identified which differ in their relationship with well-being and mental health (Litwin, 1998; Fiori et al., 2006; Litwin & Shiovitz-Ezra, 2011; Li & Zhang, 2015). Another way to determine network types is to use only characteristics of the social network as criterion variables and control for family status separately. Litwin & Stoeckel (2014) use size, composition and relational dynamic measures of the network and identify six networks (Spouse, Children, Spouse and Children, Other Family, Friend, and Other). They show that the network types are related differently to quality of life and that the frequency of the network types differs across European countries. We will follow a different approach, using only network composition to determine network types (i.e., the relative relevance of spouses, children, friends and others). We calculate network types for each country separately, taking cultural differences in network compositions into account. We chose this approach, because we want our network types to be directly linked to the family background variables whose associations with well-being we are interested in, and we control separately for contact frequency, emotional closeness and geographical proximity. In this way, we aim to identify how, for example, a Children network relates to well-being compared to a Friends network, conditional on controlling for frequency and closeness etc.

We classify the possible relationships into five categories: (1) Partner, (2) Children, (3) Other Relatives, (4) Friends, and (5) Others. Each of these categories comprises all types of relationships related to the category itself, i.e. the category Partner also includes the relationship "mother/father in law". The *relationship share* of each category in the network of a respondent is measured by the sum of the occurrence of the category divided by the network size. For each respondent, the relationship shares of all relationship categories sum to one. We use the relationship share to determine country-specific support network types according to the main source of social support. The respondents who indicate that there is no person with whom they discuss important matters are excluded. For the remaining respondents, we use hierarchical clustering with the Ward (1963) method to determine clusters which are similar with respect to the relationship shares. We choose to cut at six clusters and label them Partner, Children, Other Relatives, Family, Friends, and Diverse network. Using five clusters would not allow us to distinguish between the Friends and the Diverse network. Using more than six clusters does not provide an additional distinct network type for all countries for the five relationship categories used.

For each country, a cluster is labeled as Partner, Children, Other Relatives, Friends or Diverse network if the mean of the relationship share (averaged over all people in the cluster) of the category Partner, Children, Other Relatives, Friends and Other is higher in that cluster than in all other clusters, respectively. The labeling of the clusters would mostly be unaffected if it were instead determined by the highest mean relationship share (averaged over all people in the cluster) within a cluster, i.e., comparing across relationship type. Additionally, we include a cluster for Family networks. The Family network is the cluster with the highest sum of Partner share plus Children share plus Other Relatives share, excluding the clusters which are defined as Partner, Children, or Other Relatives network.

Apart from size and composition, a network is also characterized by relational dynamics such as geographical proximity, contact frequency and interpersonal closeness. Frequent contact with one's children appears to be associated with less depressive symptoms, albeit irrelevant of geographical proximity (Buber & Engelhardt, 2008). Closeness with the support network member affects the quality of the relationship. The number of close network members with frequent contact is positively related to less depressive symptoms (Oxman & Hull, 1997). Especially elderly people rely on members of their immediate family (partner and children) during illness (Shanas, 1979). SHARE provides different questions for these relational dynamics, which we use as controls in our analysis.

For contact frequency, the respondent is asked about the amount of contact with each person in his social support network over the last 12 months. The possible answers are (1) daily, (2) several times a week, (3) about once a week, (4) about every two weeks, (5) about once a month, (6) less than once a month, and (7) never. We recode such that the measure ranges from 0 (never) to 6 (daily). As an overall measure of the amount of network contact of a respondent, we take the average over the answers for each person in his network and call it *contact index.* E.g., if the result is 6, it means that the respondent has daily contact with all persons in his network. If it is less than 6, he must have less than daily contact with at least some member of the network.

Similar measures are constructed for proximity and closeness. The respondent is asked how far the person lives and how close he feels to the person. The categories for closeness are (0) not very close (1) somewhat close (2) very close (3) extremely close; and for proximity (0) more than 500km, (1) 100km to 500km, (2) 25km to 100km, (3) 5km to 25km, (4) 1km to 5km, and (5) less than 1km. The averages over the respective answers for each person in the respondent's network are the *closeness index* and *proximity index.* Information on the correlation of marriage, the number of children, social network dimensions and well-being measures is given in Table A5.3. We observe that the correlations between features of the family status (e.g., married) and the respective network is positive but far from perfect. That is, both people with and without children may indicate that their social support network may predominantly consist of their partner (and similarly for the other network types).

5.3 Results

5.3.1 Association of Marital Status and Parenthood with Well-Being and Mental Health

We present results in an aggregated way to illustrate the relevant patterns, and the robustness of the results with regards to confounding factors. Table 5.1 shows the associations of the three dimensions of well-being and mental health with family status (number of children, number of resident children, number of grandchildren and marital status) for all respondents (Panel I), male respondents (Panel II), and female respondents (Panel III), over all countries including country fixed effects (further country specific analyses are reported below). The table shows the raw means for each well-being measure conditional on each explanatory variable. Comparing the raw mean values for the well-being measures gives an impression of the effects sizes of each explanatory variable. However, we indicate the significance of each comparison based on regression analyses of the dependent measure on the explanatory variables; the excluded category in the regression analyses is indicated in italics in the table. We show the significance level of the variable and the direction of the association, for the regressions including controls A and B, respectively. For each set of analyses, we also indicate the sample size of the raw means, which varies across analyses because of the variation in the number of respondents in the different modules of the SHARE surveys. We use ordinary least squares for all four measures for its ease of interpretation. Detailed results for each regression are in Tables A5.4 to A5.6.

Overall we observe that marriage is consistently positively correlated with well-being and lack of depressive symptoms, which already provides evidence in favor of hypothesis i) from the introduction. We find that children are positively correlated with well-being and lack of depressive symptoms. However, our analyses show that this overall positive association is due to children after they left home: we find negative effects for the number of resident children. This pattern is consistent with the prediction of hypothesis ii) concerning the effect of non-resident children. Grandchildren correlate positively with life satisfaction and network satisfaction, but negatively with quality of life and lack of depressive symptoms, which gives us a mixed picture on the overall role of grandchildren compared to the prediction in hypothesis ii). While there are some differences in specific correlations, the overall picture is very similar for male and female respondents. Controlling for differences between the countries by conducting separate regressions for each country individually also does not qualitatively change the results (see Table A5.7). However, there is clearly heterogeneity in the sense that we observe many null effects next to those effects replicating the overall effects shown in Table 5.1.

Taken together, the results of Table 5.1 confirm findings of previous studies (Mastekaasa, 1994; Buber & Engelhardt, 2008; Bures et al., 2009; Margolis & Myrskylä, 2011; Hank & Wagner, 2013; Gibney et al., 2017) for the current large multi-country SHARE data set. Focus on the age cohort of people 50 years old and older allows us to identify different associations for children at home, children who left home already, and grandchildren. Given the consistency with previously observed patterns for the direct family status measures, we have a solid foundation for studying the broader role of family in through social networks corresponding to the different family background measures.

5.3.2 Distribution of Social Network Types

Table 5.2 presents the means of the network size, the composition measures, and relational dynamic measure of each category. For the distribution of the network size for each network type see Figure A5.3. The Partner network is a rather distinct type. It consists only of the

Table 5.1: Regressing well-being and mental health on family status for all countries.

	(I) Life satisfaction			(II) Quality of life (CASP-12)			(III) Network satisfaction			(IV) Lack of depressive symptoms (EURO-D)		
	N	Mean		N	Mean		M	Mean		N	Mean	
Panel I: All respondents												
Marriage												
Not married	15548	7.14		14902	6.67		15667	8.69		15477	7.50	
Married	36700	7.75	***A,B,+	35610	7.10	***A,B,+	36846	8.90	***A,B,+	36464	8.01	***A,+
Children												
No	4746	7.39		4569	6.92		4782	8.53		4731	7.81	
1	9613	7.37		9261	6.84	*A,+	9674	8.82	***A,B,+	9567	7.70	
2	21574	7.64	***A,B,+	20938	7.05	***A,B,+	21676	8.87	***A,B,+	21466	7.97	***A,**B,+
3 or more	16315	7.64	**A,+	15744	6.96	***A,+	16381	8.89	***A,B,+	16177	7.82	
Resident children			*A,***B,-			***A,B,-			*B,-			*B,-
Grandchildren			**B,+			***A,-			***A,B,+			***A,-
Panel II: Male respondents												
Marriage												
Not married	4576	7.17		4399	6.89		4610	8.42		4556	8.01	
Married	18271	7.78	***A,B,+	17750	7.15	***A,**B,+	18352	8.86	***A,B,+	18149	8.34	***A,+
Children												
No	2191	7.37		2114	6.99	2211	8.47		2186	8.15		
1	3885	7.49		3753	6.97	**A,+	3902	8.78	***A,*B,+	3855	8.18	
2	9595	7.73	***A,*B,+	9341	7.18	***A,**B,+	9642	8.81	***A,*B,+	9554	8.37	***A,*B,+
3 or more	7176	7.74	***A,+	6941	7.10	***A,+	7207	8.81	**A,+	7110	8.23	
Resident children						***A,B,-						
Grandchildren			**B,+			***A,-			***A,B,+			**A,-
Panel III: Female respondents												
Marriage												
Not married	10972	7.12		10503	6.58		11057	8.80		10921	7.29	
Married	18429	7.72	***A,B,+	17860	7.04	***A,B,+	18494	8.94	***A,B,+	18315	7.69	***A,+
Children												
No	2555	7.40		2455	6.85		2571	8.58		2545	7.53	
1	5728	7.29		5508	6.75		5772	8.85	***A,B,+	5712	7.39	*A,-
2	11979	7.57	**A,B,+	11597	6.95	**A,*B,+	12034	8.92	***A,B,+	11912	7.65	*B,+
3 or more	9139	7.56		8803	6.85		9174	8.95	***A,B,+	9067	7.50	
Resident children			**A,***B,-			***A,B,-						
Grandchildren						***A,-			***A,B,+			***A,-

Notes: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, +(-) indicates positive (negative) significant effect with the well-being and mental health measure; (I)-(IV) OLS Regression. Controls A: female, age, age2, country dummy; Controls B: Controls A, divorced, widowed, education, urban character of residence, employment, self-employment, health status, medication for depressive symptoms, average monthly household income dummy for low, middle, upper middle, and high income (based on country-specific 25th, 50th and 75th percentile of the average monthly household income reported in wave 2); Children: Dummy variable for having no children (excluded category), one child, two children, and three or more children. Resident children: Number of children living with their parents. If a respondent has no children then the value is set to 0; Grandchildren: Number of grandchildren, Married: Dummy variable if respondent is married or in registered partnership. Excluded category: Control A: Married but living separated from a spouse, never married, divorced, widowed, Control B: never married since a dummy variable for divorced and widowed is included in Control B. N indicates number of observations in each category for categorical variables.

partner, i.e. has a size of one, and on average has a contact index close to the maximum. Respondents which are associated with a partner network typically feel extremely or very close with their partner, resulting in the highest closeness index of all network types.

Table 5.2: Network characteristics by social network types.

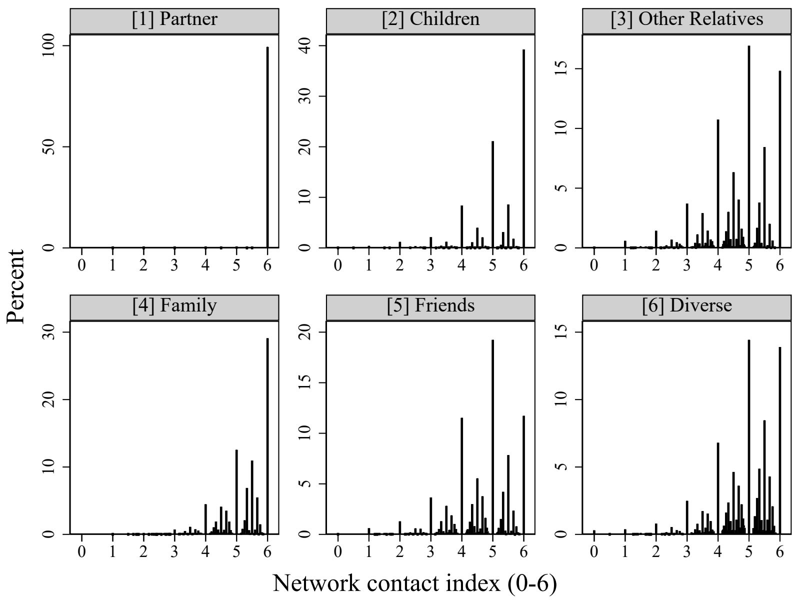
	All (1)	Partner (C1)	Other relatives (C2)	Family (C3)	Friends (C4)	Diverse (C5)	(C6)
Network size	2.60	1.00	1.93	2.88	3.21	2.95	3.53
Relationship share							
Partner	34%	100%	7%	18%	31%	14%	14%
Children	33%	0%	93%	11%	55%	10%	27%
Other Relatives	13%	0%	0%	58%	7%	8%	12%
Friends	16%	0%	0%	11%	7%	64%	12%
Others	6%	0%	0%	2%	1%	3%	34%
Contact index (0-6)	5.13	5.99	5.18	4.67	5.21	4.64	4.83
Closeness index (0-3)	2.25	2.50	2.35	2.12	2.38	1.99	2.04
Proximity index (0-5)	3.99	4.98	3.69	3.51	3.90	3.74	3.86
# obs.	50869	9254	6208	6894	13432	8498	6583
% obs.[a]	100%	18%	12%	14%	26%	17%	13%

Notes: Column (1) reports the percentages or means of respondents who have a social network and columns (C1)-(C6) for respondents associated with the respective network type. Network size: number of persons mentioned by the respondent. Relationship categories: Partner, Children, Other relatives, Friends, Other. Contact categories: (0) Never, (1) Less than once a month, (2) About once a month, (3) About every two weeks, (4) About once a week, (5) Several times a week, and (6) Daily. Closeness categories: (0) Not very close (1) Somewhat close (2) Very close (3) Extremely close. Proximity categories: (0) More than 500km, (1) 100km to 500km, (2) 25km to 100km, (3) 5km to 25km, (4) 1km to 5km, and (5) Less than 1km. Contact (closeness, proximity) index: it is defined for each respondent and is the average of the respective measure over all persons in his social support network. Relationship (contact, closeness, proximity) share: it is defined for each category of the measure and each respondent and is the sum of occurrence of each category divided by the size. Values from the same dimension may not add to 100% due to rounding.
[a] 1644 respondents report that they do not have a social network (see S4 Table).

The share of children in the network is highest for the Children network and the Family network. In both networks the second main source of support is the partner, emphasizing the importance of marital status. The Children and Family network differ in terms of the means of the relationship indices and network size, while the average contact and closeness shares are quite similar. The Other Relatives, Friends, and Diverse network types appear to be similar in terms of average contact, closeness and proximity.

We have shown that network size, as well as contact, proximity, and closeness indices differ across network types on average (see Table 5.2). Figure 5.2 presents the full distribution of the contact index for each network type (for the corresponding graphical representations of the proximity and closeness indices, see Figure A5.4 and A5.5). Even network types which are rather similar with respect to the average contact index such as the children and family networks, are rather different with respect to the individual distribution of the network contact index. This means that the actual composition of the different contact categories in the individual social network is different for each network type. Furthermore, the impact of contact, closeness, and proximity could differ, depending on the composition of a given network type: Higher self-reported closeness might, for example, denote a different depth of emotional connection in a Partner, compared to a Friends network. We therefore include the contact, closeness, and proximity indices as individual controls.

Figure 5.2: Distribution of the contact index by network type.



Notes: Each value of the network contact index is represented by a line. The height of each line represents the percentage of the index having the respective value for a network type.

Table 5.3 presents the means of gender, age and family status and the number of observations for the different network types. There are 3% of the respondents who report to have no network. The No network type and Children network type are associated with the lowest share of respondents who are married. Most respondents who have no children are associated with the Other Relatives, the Friends, or the No network type. For the means of further demographic variables see Table A5.12.

5.3.3 Association of Social Networks with Well-Being and Mental Health

We can now turn to the relationship between network characteristics and well-being and mental health measures. Table 5.4 compares well-being and mental health for respondents who have a social support network, and those who have no network at all. We use the different network types as explanatory variable, and include network size, family status variables (as in Table 5.1), and socioeconomic variables as controls. We do not control for relational dynamics, because these measures are not defined for those who have no social support network. For the detailed regression results see Tables A5.8 to A5.10 and for the raw means conditional on network size

Table 5.3: Family status by social network types.

Network types	No Network C0	Partner C1	Children C2	Other relatives C3	Family C4	Friends C5	Diverse C6
Female	49%	35%	72%	61%	56%	61%	63%
Age at interview	67	65	71	63	66	64	65
Marital Status							
Married/registered partnership	50%	94%	41%	62%	86%	59%	60%
Divorced/living separated	15%	3%	12%	12%	6%	16%	14%
Widowed	23%	1%	45%	12%	7%	15%	19%
Parenthood							
Number of children	1.97	2.24	2.55	1.76	2.39	1.82	2.06
Number of resident children	0.29	0.36	0.29	0.33	0.33	0.31	0.35
Number of grandchildren	2.59	2.60	3.89	1.85	2.97	1.91	2.36
Having grandchildren	64%	69%	87%	53%	79%	57%	66%
% obs.	3%	18%	12%	13%	26%	16%	13%

Notes: Columns (C0)-(C6) report the percentages or means of respondents associated with the respective network type. There are in total 52513 observations. Values from the same dimension may not add to 100% due to rounding.

see Table A5.11.

In accordance with hypothesis iii), all network types relate positively to measures of well-being, for both males and females, even after controlling for family structure. The effect is consistently observed for Life satisfaction and Network satisfaction. For CASP-12 the effect is observed for Partner and Friends network types for male respondents and for Partner, Family and Friends network types for female respondents. Interestingly, the positive relationship with Children network and Lack of depressive symptoms mostly emerges only after inclusion of the full set of controls. Network size is positively related to all measures of well-being (Tomini et al., 2016). The results obtained in this network analysis support the broader relevance of family through the resulting networks for well-being and mental health as postulated in hypotheses i) and ii).

We next compare the different network types with each other, accounting for variation of the relational indices across the different network types. Table 5.5 shows the results confined to respondents who indicated the presence of some social support network. For the detailed regression results see Tables A5.12 to A5.14. The excluded category of the network type is the Partner network, which had consistently strong and significant associations in Table 5.4, and is taken as a benchmark here.

We find that for Life satisfaction, CASP-12, and Lack of depressive symptoms, the more diverse networks have typically weaker associations than the Partner network, with the exception of the Friends network for CASP-12. In contrast, Network satisfaction is consistently higher for all other networks, except for the Diverse network. Note, that this emerges despite controlling for network size, contact, closeness and proximity index. Fiori et al. (2006) pointed out that support quality is an important factor for depressive symptoms. We find consistently that the closeness and contact measure is positively correlated with mental health and well-being. However, we observe a negative relationship of mere proximity with well-being and mental health. While for

Table 5.4: Regressing well-being and mental health on social support network types; controlling for network size and family status for all countries.

Dependent Variable	(I) Life satisfaction			(II) Quality of life (CASP-12)			(III) Network satisfaction			(IV) Lack of depressive symptoms (EURO-D)		
	N	Mean		N	Mean		M	Mean		N	Mean	
Panel I: All respondents												
Network type												
No network	1624	6.93		1555	6.48		1644	6.49		1615	7.59	
Partner	9209	7.7	***A,B,+	8944	7.05	***A,B,+	9254	8.98	***A,B,+	9145	8.21	***A,B,+
Children	6157	7.37	***A,B,+	5909	6.52		6208	9.10	***A,B,+	6116	7.49	***B,+
Other Relatives	6856	7.44	***A,B,+	6622	7.06	*B,+	6894	8.77	***A,B,+	6805	7.76	
Family	13381	7.77	***A,B,+	12989	7.05	*B,+	13432	9.05	***A,B,+	13316	7.98	*A,***B,+
Friends	8468	7.52	***A,B,+	8200	7.13	**B,+	8498	8.73	***A,B,+	8423	7.82	
Diverse	6553	7.5	***B,+	6293	6.92		6583	8.74	***A,B,+	6521	7.68	*A,-
Network size			***A,B,+			***A,B,+			***A,B,+			***A,**B,+
Panel II: Male respondents												
Network type												
No network	829	6.98		790	6.67		841	6.60		823	7.99	
Partner	5996	7.70	***A,B,+	5820	7.06	*A,B,+	6026	9.03	***A,B,+	5955	8.36	**A,B,+
Children	1755	7.77	***A,B,+	1695	7.03		1765	9.06	***A,B,+	1745	8.17	*B,+
Other Relatives	2656	7.47	***A,B,+	2576	7.15		2671	8.65	***A,B,+	2634	8.18	
Family	5895	7.77	***A,B,+	5732	7.07		5922	9.00	***A,B,+	5870	8.31	**B,+
Friends	3310	7.64	***A,B,+	3218	7.27	*A,B,+	3322	8.58	***A,B,+	3286	8.25	
Diverse	2406	7.65	**A,B,+	2318	7.16		2415	8.54	***A,B,+	2392	8.23	
Network size			***A,B,+			***A,B,+			***A,B,+			***A,+
Panel III: Female respondents												
Network type												
No network	795	6.88		765	6.29		803	6.38		792	7.18	
Partner	3213	7.71	***A,B,+	3124	7.03	***A,B,+	3228	8.90	***A,B,+	3190	7.93	***A,B,+
Children	4402	7.21	**B,+	4214	6.32		4443	9.12	***A,B,+	4371	7.22	**B,+
Other Relatives	4200	7.43	**B,+	4046	6.99		4223	8.84	***A,B,+	4171	7.49	
Family	7486	7.77	**A,***B,+	7257	7.04	*B,+	7510	9.10	***A,B,+	7446	7.72	***B,+
Friends	5158	7.45	*B,+	4982	7.05	*B,+	5176	8.82	***A,B,+	5137	7.55	
Diverse	4147	7.41	**B,+	3975	6.79		4168	8.86	***A,B,+	4129	7.36	
Network size			***A,B,+			***A,B,+			***A,B,+			***A,*B,+

Notes: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, +(-) indicates positive (negative) significant effect with the well-being measure; (I)-(IV) OLS Regression. Controls A: female, age, age2, country dummy; Controls B: Controls A, divorced, widowed, education, household size, urban character of residence, retired, self-employment, health status, average monthly household income dummy for low, middle, upper middle, and high income (based on country-specific 25th, 50th and 75th percentile of the average monthly household income reported in wave 2). Children, resident children, grandchildren and married included in both A and B. Network types: the excluded category is having no network. N indicates number of observations in each category for categorical variables.

Table 5.5: Regressing well-being and mental health on social support network types (only for respondents with a social network); controlling for network size, relational dynamic measures, and family status for all countries.

Dependent Variable	(I) Life satisfaction	(II) Quality of life	(III) Network satisfaction (CASP-12)	(IV) Lack of depressive symptoms (EURO-D)
Panel I: All respondents				
Network type[a]				
Children	***A,B,-	***A,B,-	***A,B,+	***A,*B,-
Other Relatives	***A,-		***A,B,+	***A,B,-
Family	***A,-	***A,-	***A,B,+	***A,-
Friends		***B,+	***A,B,+	***A,B,-
Diverse	***A,-	***A,-		***A,B,-
Network size	***A,B,+	***A,B,+	***A,B,+	***A,B,+
Contact Index	***A,B,+	***A,B,+	***A,B,+	**A,B,+
Closeness Index	***A,B,+	***A,B,+	***A,B,+	***A,B,+
Proximity Index	***A,-	***A,B,-	***A,B,-	**A,-
Panel II: Male respondents[b]				
Network type[a]				
Children	*A,-	*A,-	***A,B,+	*A,-
Other Relatives		*B,+		*A,-
Family		*A,-		
Friends		***A,B,+		
Diverse				*A,-
Network size	***A,B,+	***A,B,+	***A,B,+	***A,*B,+
Contact Index	**A,B,+	**A,***B,+	***A,B,+	
Closeness Index	***A,B,+	***A,B,+	***A,B,+	***A,B,+
Proximity Index	**A,-	***A,-	***A,B,-	
Panel III: Female respondents[b]				
Network type[a]				
Children	***A,B,-	***A,B,-	***A,B,+	***A,B,-
Other Relatives	***A,*B,-	***A,-	***A,B,+	***A,B,-
Family	***A,-	***A,**B,-	***A,B,+	***A,**B,-
Friends	***A,*B,-		***A,B,+	***A,B,-
Diverse	***A,*B,-	***A,*B,-	***A,B,+	***A,B,-
Network size	***A,B,+ ***A,B,+	***A,B,+	***A,**B,+	
Contact Index	***A,B,+	***A,B,+	***A,B,+	*A,B,+
Closeness Index	***A,B,+	***A,B,+	***A,B,+ ***A,B,+	
Proximity Index	**A,-	***A,**B,-	***A,B,-	**A,-

Notes: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, +(-) indicates positive (negative) significant effect with the well-being measure; (I)-(IV) OLS Regression. Controls A: female, age, age2, country dummy; Controls B: Controls A, divorced, widowed, education, household size, urban character of residence, retired, self-employment, health status, average monthly household income dummy for low, middle, upper middle, and high income (based on country-specific 25th, 50th and 75th percentile of the average monthly household income reported in wave 2). Children, resident children, grandchildren and married included in both A and B.

[a]Partner network is the excluded category

[b]Control female is excluded

the associations with family status no relevant gender differences were observed, we observe that associations with network types differ for male and female respondents. For male respondents in most cases the effects of the network types are not significantly different from the Partner network. For the females, the above discussed associations show up significantly.

5.4 Discussion

Reflecting the results presented in the previous section in terms of our three research hypotheses, the following implications for the well-being and mental health of people aged 50 and above emerge: i) There appears to be a strong positive association of being married/having a partner as part of a social network. ii) Non-residential children also relate positively to well-being and mental health. On the other hand, the effect of grandchildren in general appears to be mixed. While they may be associated with higher life and network satisfaction, the same does not appear to hold for depressive symptoms and perceived quality of life. iii) We find clear evidence of positive relationships of all types of social networks with our measures of well-being, over and beyond the respective underlying family status indicators. Hence, a simple focus on family status measures, not accounting for the resulting network structures, misses important aspects of the relationship of family and well-being and mental health.

In contrast to negative associations reported in many studies (for an overview see Hansen 2012, or the discussion in Nelson et al. 2013 and Herbst and Ifcher 2016), we find that children are indeed positively correlated with well-being and lack of depressive symptoms, when controlling for residential status (resident children are negatively associated with well-being). This result is consistent with age-dependence in the correlation of children with well-being (Mastekaasa, 1994; Margolis & Myrskylä, 2011) and mental health (Buber & Engelhardt, 2008; Hank & Wagner, 2013). The results suggest that the finding of a negative link between children and well-being and mental health may not generalize to older people whose children have often left home already. As stress associated with balancing the competing demands of childcare, work and personal life decreases, once people get older and their children leave house, the importance of children as caregivers and social contacts might prevail. The mixed effect of grandchildren is more difficult to explain. Potentially, there are positive effects of having grandchildren in terms of social support that might coincide with negative aspects, such as having to care for these grandchildren (Gerard et al., 2006; Leder et al., 2007). As the SHARE data set only provides us with rudimentary information about grandchildren (there is for example no information about the residency of them), we cannot shed more light on the relation between grandchildren and well-being, as well as mental health.

We observe that all types of networks have positive associations with our dependent measures. Network characteristics such as size, closeness, contact frequency and proximity are also relevant indicators of well-being and mental health. For male respondents in most cases the effects of the network types are not significantly different from the Partner network. For female respondents, on the other hand, we observe more cases where associations of well-being and mental health with the Partner network type are significantly different from those for the other network types. Overall we find that especially the Partner network is consistently positively correlated with well-being and mental health, despite the small network size of 1. This is in contrast to Litwin and Stoeckel (2013; 2014), who found that the Spouse network is not significantly related to well-being. However, importantly, because we control for network size separately, positive associations with size are captured by this variable. A remarkable feature of the findings in Table 5.4 is that network characteristics are positively associated with well-being and mental health even after controlling for the above-shown associations with family status indicators. That is, a healthy partner network captures more than just being married, as do other types of networks. This fits previous results, suggesting that it is not being married per se, but being satisfied with the relationship that is associated with less depressive symptoms (Gove et al., 1983; Hank & Wagner, 2013). Kim and McKenry also report both, a positive relationship of well-being with being married, and an additional role for the perceived quality of the marriage on top of that (Kim & McKenry, 2002). Our research extends these previous studies, by demonstrating the role of both the presence of a partner and the associated network, where the partner actively provides social support. As the size of a social network seems to be an important driver of subjective well-being (Wang, 2016), this could indicate that a small partner network can offset the lack of a larger social network. Unfortunately, a limitation of our present study is that besides general network satisfaction, the SHARE data set has no more fine-grained questions for the quality of marriages/partnerships.

Taken together, our results suggest that social networks may be important for well-being and mental health in old age. Spouses, partners and children are often the basis of long-lasting social networks, which can provide social support to elderly people. However, different forms of network may have similar effects, as our data especially for male respondents suggests. As discussed above, this might derive from a level of trust and reciprocity implicit in all forms of networks. A remaining limitation of our study is of course that the results are correlational in nature. Further studies, comparing for example well-being and mental health before and after the formation of partnerships or social networks in longitudinal data are needed to establish which factors cause the positive effects found here and in the literature. Furthermore, research suggests that there is an important link between social support obtained from social networks and subjective well-being (Siedlecki et al., 2014; Wang, 2016). Subjective well-being is commonly measured with questions concerning life satisfaction, positive affect (experiencing positive emotions), and negative affect (experiencing negative emotions). As the SHARE data set only measures life satisfaction, we cannot draw a complete picture of the effects of family status and social networks on broader measures of subjective well-being.

Networks may exert an influence on the person's life beyond the mere role of the corresponding family status, for example by moderating influences of the environment on well-being. The direct association of family status with well-being and mental health may not capture such effects. Importantly, the current insights need to guide further research, with the next step being the assessment of the causal direction of the reported associations. This will allow moving towards making recommendations for public policy to maintain the well-being and mental health of the elderly through social networks.

Chapter 5 References

Beckie, T. M., & Hayduk, L. A. (1997). Measuring Quality of Life. *Social Indicators Research*, *42*, 21–39.

Blieszner, R., & Mancini, J. A. (1987, April). Enduring Ties: Older Adults' Parental Role and Responsibilities. *Family Relations*, *36*(2), 176.

Brandt, M., Haberkern, K., & Szydlik, M. (2009, October). Intergenerational Help and Care in Europe. *European Sociological Review*, *25*(5), 585–601.

Buber, I., & Engelhardt, H. (2008, March). Children's impact on the mental health of their older mothers and fathers: findings from the Survey of Health, Ageing and Retirement in Europe. *European Journal of Ageing*, *5*(1), 31–45.

Bures, R. M., Koropeckyj-Cox, T., & Loree, M. (2009, May). Childlessness, parenthood, and depressive symptoms among middle-aged and older adults. *Journal of Family Issues*, *30*(5), 670–687.

Börsch-Supan, A. (2018a). Survey of health, ageing and retirement in europe (share) wave 1. release version: 7.0.0. *SHARE-ERIC*.

Börsch-Supan, A. (2018b). Survey of health, ageing and retirement in europe (share) wave 2. release version: 70.0. *SHARE-ERIC*.

Börsch-Supan, A. (2018c). Survey of health, ageing and retirement in europe (share) wave 4. release version: 7.0.0. *SHARE-ERIC*.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., ... Zuber, S. (2013, August). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, *42*(4), 992–1001.

Börsch-Supan, A., Brandt, M., Litwin, H., & Weber, G. E. (2015). *Active ageing and solidarity between generations in Europe: first results from SHARE after the economic crisis 2013.* De Gruyter: Berlin.

Castro-Costa, E., Dewey, M., Stewart, R., Banerjee, S., Huppert, F., Mendonca-Lima, C., ... Prince, M. (2008, March). Ascertaining late-life depressive symptoms in Europe: an evaluation of the survey version of the EURO-D scale in 10 nations. The SHARE project. *International Journal of Methods in Psychiatric Research*, *17*(1), 12–29.

Cobb, S. (1976). Social support as a moderator of life stress. *Psychosomatic medicine*, *38*(5), 300–314. (Publisher: LWW)

Definition of social support - NCI Dictionary of Cancer Terms - National Cancer Institute. (2023, October). Retrieved 2023-03-17, from `https://web.archive.org/web/`

`20181023124317/https://www.cancer.gov/publications/dictionaries/cancer-terms/def/social-support`

Diener, E., & Suh, E. (1997). Measuring Quality of Life: Economic, Social, and Subjective Indicators. *Social Indicators Research*, *40*, 189–216.

Dodge, R., Daly, A., Huyton, J., & Sanders, L. (2012). The challenge of defining wellbeing. *International Journal of Wellbeing*, *2*(3), 222–235.

Evans, J. (2011). Findings from the National Well-being Debate. *London: Office for National Statistics*.

Evenson, R. J., & Simon, R. W. (2005). Clarifying the Relationship between Parenthood and Depression. *Journal of Health and Social Behavior*, *46*(4), 341–358.

Fiori, K. L., Antonucci, T. C., & Cortina, K. S. (2006, January). Social Network Typologies and Mental Health Among Older Adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *61*(1), 25–32.

Fung, H. H., Carstensen, L. L., & Lang, F. R. (2001). Age-Related Patterns in Social Networks among European Americans and African Americans: Implications for Socioemotional Selectivity across the Life Span. *The International Journal of Aging and Human Development*, *52*(3), 185–206.

Gerard, J. M., Landry-Meyer, L., & Roe, J. G. (2006). Grandparents raising grandchildren: The role of social support in coping with caregiving challenges. *The International Journal of Aging and Human Development*, *62*(4), 359–383.

Gibney, S., Delaney, L., Codd, M., & Fahey, T. (2017, January). Lifetime Childlessness, Depressive Mood and Quality of Life Among Older Europeans. *Social Indicators Research*, *130*(1), 305–323.

Gove, W. R., Hughes, M., & Style, C. B. (1983). Does marriage have positive effects on the psychological well-being of the individual? *Journal of health and social behavior*, 122–131.

Hank, K., & Wagner, M. (2013, November). Parenthood, Marital Status, and Well-Being in Later Life: Evidence from SHARE. *Social Indicators Research*, *114*(2), 639–653.

Hansen, T. (2012). Parenthood and Happiness: a Review of Folk Theories Versus Empirical Evidence. *Social Indicators Research*, *108*(1), 29–64.

Hansen, T., Slagsvold, B., & Moum, T. (2009). Childlessness and Psychological Well-Being in Midlife and Old Age: An Examination of Parental Status Effects across a Range of Outcomes. *Social Indicators Research*, *94*(2), 343–362.

Herbst, C. M., & Ifcher, J. (2016, September). The increasing happiness of US parents. *Review of Economics of the Household*, *14*(3), 529–551.

Hyde, M., Wiggins, R. D., Higgs, P., & Blane, D. B. (2003, May). A measure of quality of life in early old age: The theory, development and properties of a needs satisfaction model (CASP-19). *Aging & Mental Health*, *7*(3), 186–194.

International Standard Classification of Education, ISCED 1997. (2003). In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables* (pp. 195–220). Boston, MA: Springer US.

Kim, H. K., & McKenry, P. C. (2002). The relationship between marriage and psychological well-being: A longitudinal analysis. *Journal of family Issues*, *23*(8), 885–911.

Knesebeck, O. V. D., Wahrendorf, M., Hyde, M., & Siegrist, J. (2007, March). Socio-economic position and quality of life among older people in 10 European countries: results of the SHARE study. *Ageing and Society*, *27*(02), 269–284.

Leder, S., Grinstead, L. N., & Torres, E. (2007). Grandparents raising grandchildren: Stressors, social support, and health outcomes. *Journal of Family Nursing*, *13*(3), 333–352.

Li, T., & Zhang, Y. (2015, April). Social network types and the health of older adults: Exploring reciprocal associations. *Social Science & Medicine*, *130*, 59–68.

Litwin, H. (1998, February). Social network type and health status in a national sample of elderly Israelis. *Social Science & Medicine*, *46*(4-5), 599–609.

Litwin, H. (2010, September). Social Networks and Well-being: A Comparison of Older People in Mediterranean and Non-Mediterranean Countries. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *65B*(5), 599–608.

Litwin, H., & Shiovitz-Ezra, S. (2011, June). Social Network Type and Subjective Well-being in a National Sample of Older Americans. *The Gerontologist*, *51*(3), 379–388.

Litwin, H., & Stoeckel, K. J. (2013, October). Social networks and subjective wellbeing among older Europeans: does age make a difference? *Ageing and Society*, *33*(07), 1263–1281.

Litwin, H., & Stoeckel, K. J. (2014, October). Confidant Network Types and Well-Being Among Older Europeans. *The Gerontologist*, *54*(5), 762–772.

Malter, F., & Börsch-Supan, A. E. (2013). SHARE Wave 4: Innovations & Methodology. *Munich: MEA, Max Planck Institute for Social Law and Social Policy*.

Manuel, J. I., Martinson, M. L., Bledsoe-Mansori, S. E., & Bellamy, J. L. (2012, December). The influence of stress and social support on depressive symptoms in mothers with young children. *Social Science & Medicine*, *75*(11), 2013–2020.

Margolis, R., & Myrskylä, M. (2011, March). A Global Perspective on Happiness and Fertility. *Population and Development Review*, *37*(1), 29–56.

Mastekaasa, A. (1994). Marital Status, Distress, and Well-Being: An International Comparison. *Journal of Comparative Family Studies*, *25*(2), 183–205.

Matheson, J. (2011). Measuring What Matters: National Statistician's Reflections on the National Debate on Measuring National Well-being. *London: Office for National Statistics*.

Nelson, S. K., Kushlev, K., English, T., Dunn, E. W., & Lyubomirsky, S. (2013, January). In Defense of Parenthood: Children Are Associated With More Joy Than Misery. *Psychological Science*, *24*(1), 3–10.

Oxman, T. E., & Hull, J. G. (1997, January). Social Support, Depression, and Activities of Daily Living in Older Heart Surgery Patients. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *52B*(1), P1–P14.

Pavot, W., & Diener, E. (1993). The Affective and Cognitive Context of Self-Reported Measures of Subjective Well-Being. *Social Indicators Research*, *28*(1), 1–20.

Pearson, J. E. (1986). The definition and measurement of social support. *Journal of Counseling & Development*. (Publisher: American Association for Counseling and Development)

Pinquart, M., & Sörensen, S. (2000). Influences of Socioeconomic Status, Social Network, and Competence on Subjective Well-Being in Later Life: A Meta-Analysis. *Psychology and Aging*, *15*(2), 187–224.

Prince, M. J., Reischies, F., Beekman, A. T. F., Fuhrer, R., Jonker, C., Kivela, S. L., ... Copeland, J. R. M. (1999, April). Development of the EURO–D scale – a European Union initiative to compare symptoms of depression in 14 European centres. *British Journal of Psychiatry*, *174*(04), 330–338.

Shanas, E. (1979, April). The Family as a Social Support System in Old Age. *The Gerontologist*, *19*(2), 169–174.

Siedlecki, K. L., Salthouse, T. A., Oishi, S., & Jeswani, S. (2014). The relationship between social support and subjective well-being across age. *Social indicators research*, *117*(2), 561–576. (Publisher: Springer)

Tomini, F., Tomini, S. M., & Groot, W. (2016, December). Understanding the value of social networks in life satisfaction of elderly people: a comparative study of 16 European countries using SHARE data. *BMC Geriatrics*, *16*(1).

Wang, X. (2016). Subjective well-being associated with size of social network and social support of elderly. *Journal of health psychology*, *21*(6), 1037–1042.

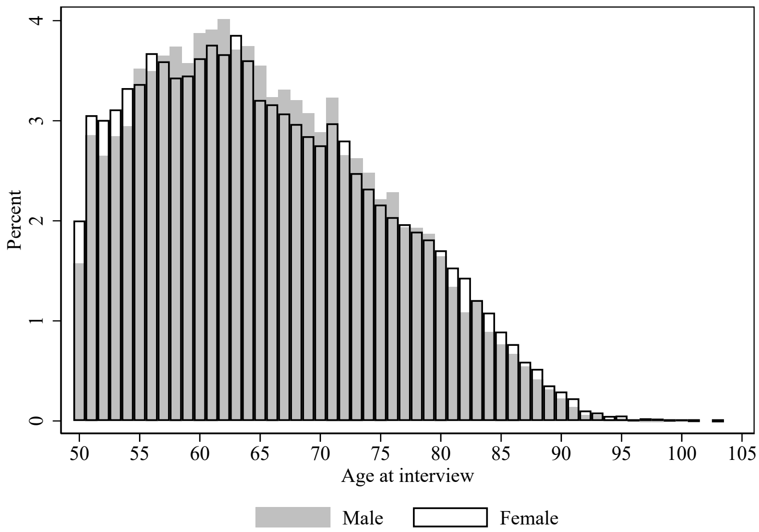
Ward, J. H. (1963, March). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

Chapter 5 Appendix

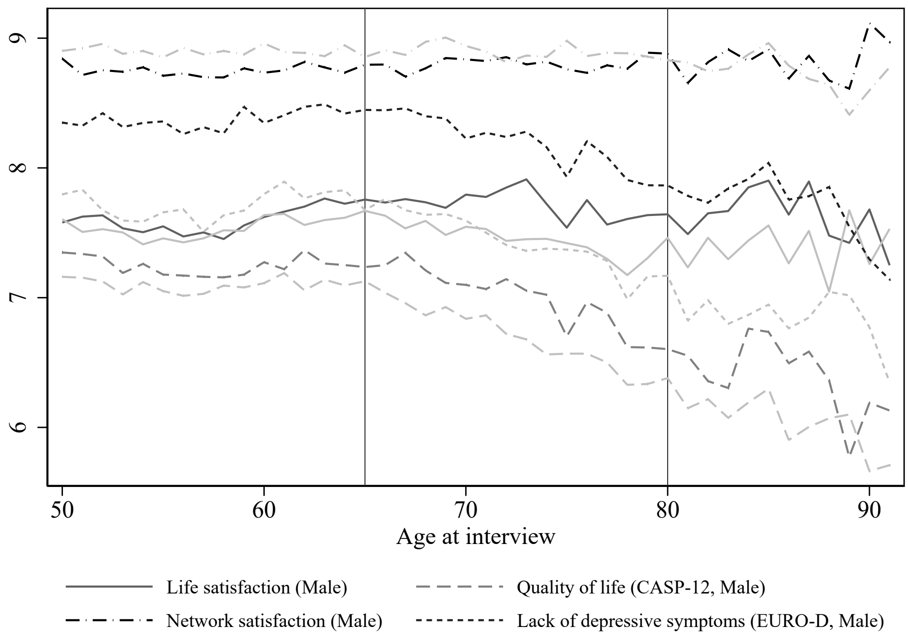
Figures

Figure A5.1: Age distribution.



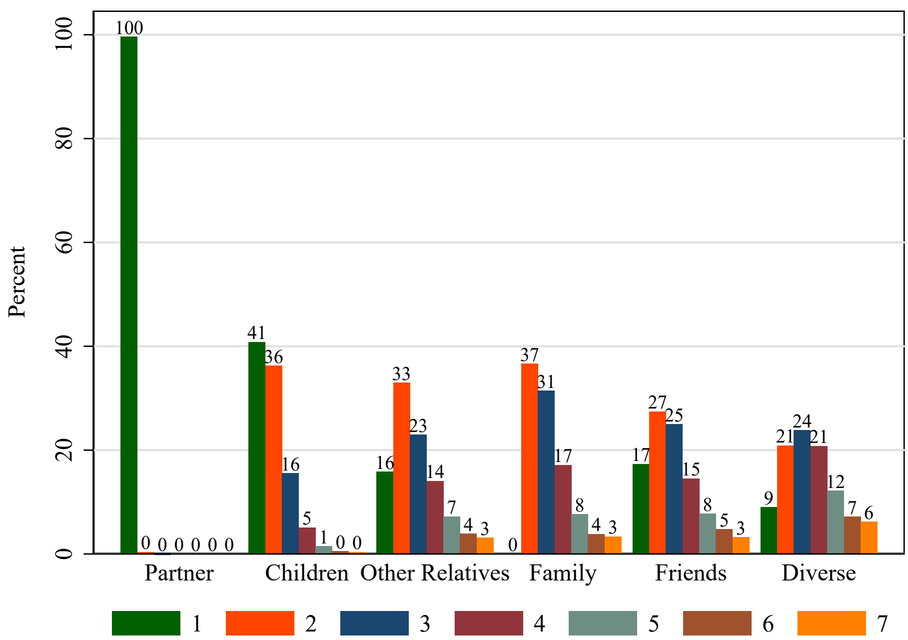
Notes: Percent of male (female) respondents for each age.

Figure A5.2: Average well-being and mental health measure, by gender.



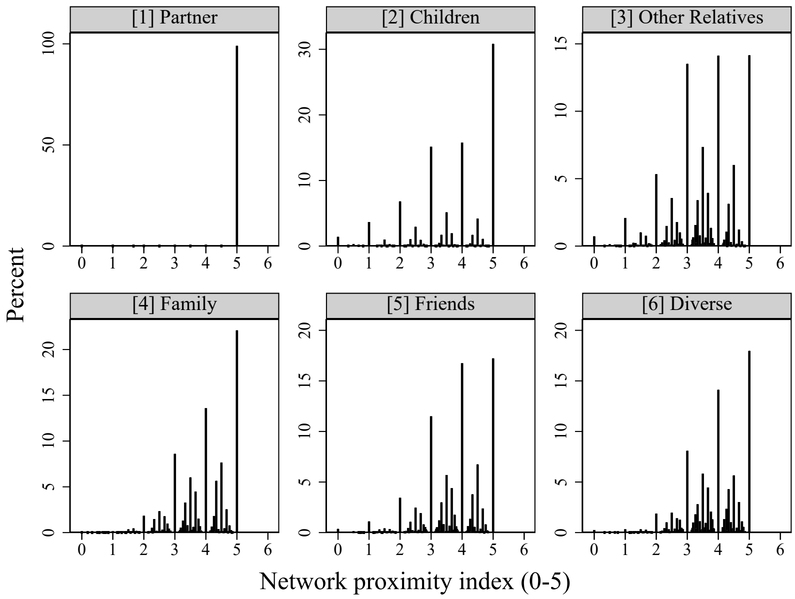
Notes: Average well-being and mental health measure for all ages from 50 to 90 years for male and female respondents. Male: black lines, Female: grey lines. After age 91 the number of available observations drops to less than 50.

Figure A5.3: Network size by network type.



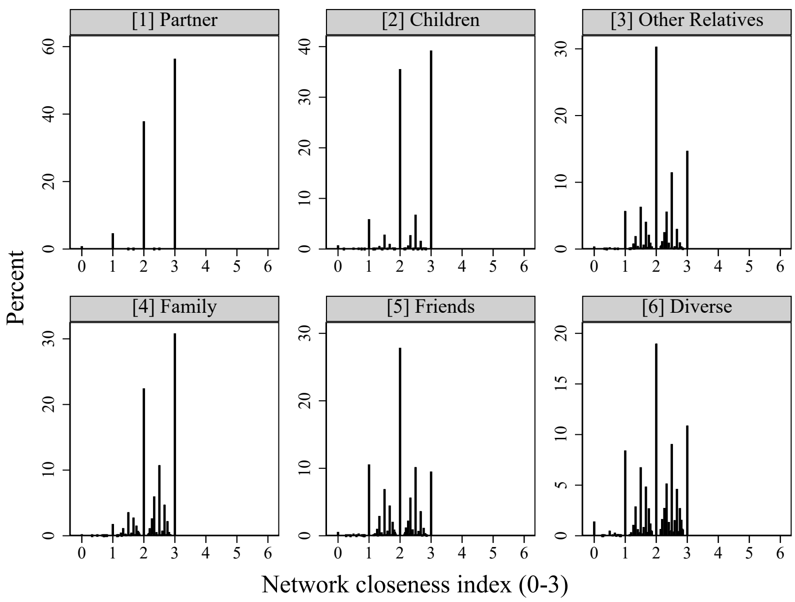
Notes: The size of a bar reflects the share of respondents in a network type having a network of size 0 to 7.

Figure A5.4: Distribution of the proximity index by network type.



Notes: Each value of the network contact index is represented by a line. The height of each line represents the percentage of the index having the respective value for a network type.

Figure A5.5: Distribution of the closeness index by network type.



Notes: Each value of the network contact index is represented by a line. The height of each line represents the percentage of the index having the respective value for a network type.

Tables

Table A5.1: Number of observations and unfolding income brackets per country.

		Unfolding brackets in €				Percent of household income			
Country	# obs.	Percent	Low	Middle	High	[0,Low)	[Low,Middle)	[Middle,High)	[High,)
Austria	4808	9%	1500	2000	3000	30%	17%	24%	29%
Belgium	4844	9%	1500	2000	3000	28%	20%	24%	28%
Czech Republic	4898	9%	1629	2037	2851	84%	5%	2%	9%
Denmark	2074	4%	1342	2013	2684	9%	15%	16%	60%
Estonia	6330	12%	192	256	320	3%	4%	12%	81%
France	5111	10%	1500	2000	3000	33%	17%	23%	28%
Germany	1429	3%	1500	2000	3000	27%	21%	25%	27%
Hungary	2901	6%	1444	2166	2888	91%	2%	1%	6%
Italy	3317	6%	1500	2000	3000	45%	18%	18%	19%
Netherlands	2498	5%	1500	2000	3000	19%	22%	29%	30%
Poland	1486	3%	1418	1890	2363	82%	1%	1%	16%
Portugal	1837	3%	1500	2000	3000	64%	10%	7%	19%
Slovenia	260	5%	150	2000	3000	69%	12%	7%	11%
Spain	3081	6%	1500	2000	3000	57%	12%	10%	21%
Sweden	1816	3%	1108	2215	3323	7%	30%	30%	32%
Switzerland	3483	7%	1620	2025	2430	10%	5%	6%	79%
Total	52513								

Notes: Low, Middle, and High are the 25^{th}, 50^{th}, and 75^{th} percentile of the reported household incomes from SHARE wave 2 (Field time 2006-2007). If a respondent refuses to state the amount of the overall income, after tax that the entire household had in an average month of the last year the interviewer asks the following questions, starting with the lowest threshold: Do you earn a) more than this amount, b) less than this amount or c) approximately this amount. The boundaries of the intervals are the respective country-specific 25^{th}, 50^{th}, and 75^{th} percentile of the reported household incomes from SHARE wave 2 which are used as unfolding brackets in SHARE wave 4.

Table A5.2: Summary statistics of demographic variables.

	All (1)	Male (2)	Female (3)	All (4)
Female	56%	0%	100%	52513
Age at interview	66	66	66	52513
Parenthood				
Number of children	2.15	2.16	2.15	52513
Having children	91%	90%	91%	52513
No children	9%	10%	9%	52513
One child	18%	17%	20%	52513
Two children	41%	42%	41%	52513
Three or more children	31%	31%	31%	52513
Number of resident children[a]	0.33	0.36	0.31	52513
Having resident children[b]	24%	24%	23%	52513
No resident child[b]	74%	73%	75%	47731
All children are resident[b]	9%	10%	8%	47731
Number of grandchildren	2.61	2.43	2.74	52513
Having grandchildren	69%	66%	71%	52513
Marital Status				
Married/registered partnership	70%	80%	63%	52513
Divorced/living separated	10%	8%	11%	52513
Widowed	14%	6%	21%	52513
Housing				
Household size	2.16	2.28	2.06	52513
Big city	14%	14%	15%	50879
Suburbs of big city	10%	11%	10%	50879
Large town	16%	16%	17%	50879
Small town	24%	24%	24%	50879
Rural area/villag	35%	36%	34%	50879
Average monthly household income				
Low income	38%	35%	41%	48736
Middle income	13%	13%	12%	48736
Upper middle income	15%	15%	15%	48736
High income	34%	37%	32%	48736
Employed	25%	26%	23%	52466
Self-employed	6%	8%	4%	52466
Education				
None	3%	2%	3%	52179
Primary school	19%	17%	21%	52179
Lower secondary school	19%	18%	21%	52179
Upper secondary school	34%	37%	32%	52179
Post-secondary non-tertiary education	5%	5%	5%	52179
First stage tertiary education	19%	21%	18%	52179
Second stage tertiary education	1%	1%	1%	52179
Health status				
Poor	12%	11%	13%	52500
Fair	30%	28%	31%	52500
Good	35%	36%	35%	52500
Very good	16%	18%	15%	52500
Excellent	7%	7%	6%	52500
Medication for depressive symptoms	13%	8%	17%	52461

Notes: Columns (1)-(3) report the percentages or means of all respondents, and by gender. Column (4) shows the total number of observations. SHARE is using the international classification of education ISCED-97 with which education can be classified according to internationally agreed set of definitions and concepts (UNESCO 1997). Medication for depressive symptoms is equal to one if the respondent takes drugs for sleeping problems, anxiety or depression. Values from the same dimension may not add to 100% due to rounding.

[a]The number of resident children is inferred from matching the age/gender information from the persons living with the family (Coverscreen module) and the age/gender information from the children of the respondents (Children module).

[b]Percentage conditional on having a child. UNESCO. 1997. "International Standard Classification of Education ISCED 1997."

Table A5.3: Correlation of family status, social network characteristics, well-being and mental health.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 Number of children															
2 Number of resident children	0.23														
3 Married/registered partnership	0.14	0.11													
4 Life satisfaction	0.03	-0.01	0.15												
5 Quality of life (CASP-12)		-0.03	0.11	0.57											
6 Network satisfaction	0.05		0.07	0.25	0.18										
7 Lack of depressive symptoms (EURO-D)		0.01	0.12	0.44	0.55	0.12									
8 Network size	0.08	0.02	0.03	0.10	0.12	0.11	0.02								
Relationship share															
9 Partner	0.05	0.05	0.47	0.08	0.06	0.05	0.12	-0.44							
10 Children	0.20	-0.01	-0.16	-0.04	-0.11	0.11	-0.08	0.19	-0.44						
11 Other relatives	-0.17		-0.16	-0.03	0.01	-0.04	-0.03	0.14	-0.30	-0.28					
12 Friends	-0.12	-0.03	-0.19		0.07	-0.09	-0.01	0.17	-0.36	-0.32	-0.10				
13 Others	-0.07	-0.03	-0.13	-0.06	-0.04	-0.11	-0.06	0.06	-0.21	-0.18	-0.06	-0.08			
14 **Contact index (0-6)**	0.06	0.13	0.21	0.01	-0.05	0.18	0.02	-0.36	0.50	0.01	-0.25	-0.35	-0.15		
15 **Closeness Index (0-3)**	0.08	0.05	0.14	0.14	0.08	0.35	0.08	-0.09	0.24	0.13	-0.10	-0.25	-0.25	0.35	
16 **Proximity Index (0-5)**	0.04	0.16	0.22		-0.07	0.03	0.02	-0.34	0.51	-0.20	-0.24	-0.19	-0.02	0.58	0.16

Notes: Correlations with a p-value smaller than 0.05 are shown. Network size: number of persons mentioned by the respondent. Relationship categories: Partner, Children, Other relatives, Friends, Other. Contact categories: (0) Never, (1) Less than once a month, (2) About once a month, (3) About every two weeks, (4) About once a week, (5) Several times a week, and (6) Daily. Closeness categories: (1) Not very close (2) Somewhat close (3) Very close (4) Extremely close. Proximity categories: (0) More than 500km, (1) 100km to 500km, (2) 25km to 100km, (3) 5km to 25km, (4) 1km to 5km, and (5) Less than 1km. Contact (closeness, proximity) index: it is defined for each respondent and is the average of the respective measure over all persons in his social support network. Relationship share: it is defined for each category of the measure and each respondent and is the sum of occurrence of each category divided by the size.

Table A5.4: Regressing well-being and mental health on family status for all countries, all respondents.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
Married/registered partnership	0.56***	0.44***	0.30***	0.20***	0.21***	0.26***	0.28***	0.039
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.321)
[1] Having 1 child	0.0046	-0.025	0.066*	0.028	0.22***	0.17***	-0.043	-0.015
	(0.891)	(0.466)	(0.035)	(0.362)	(0.000)	(0.000)	(0.206)	(0.651)
[2] Having 2 children	0.18***	0.11***	0.19***	0.11***	0.24***	0.19***	0.11***	0.095**
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.003)
[3] Having 3 or more children	0.11**	0.057	0.13***	0.051	0.23***	0.17***	0.021	0.019
	(0.001)	(0.113)	(0.000)	(0.123)	(0.000)	(0.000)	(0.555)	(0.588)
Number of resident children	-0.031*	-0.054***	-0.097***	-0.12***	-0.016	-0.023*	-0.020	-0.027*
	(0.019)	(0.000)	(0.000)	(0.000)	(0.130)	(0.047)	(0.152)	(0.033)
Number of grandchildren	-0.0050	0.010**	-0.022***	-0.00015	0.015***	0.018***	-0.020***	-0.0022
	(0.173)	(0.003)	(0.000)	(0.961)	(0.000)	(0.000)	(0.000)	(0.521)
Controls								
Female	-0.030	0.053***	-0.15***	-0.041**	0.14***	0.15***	-0.65***	-0.51***
	(0.056)	(0.001)	(0.000)	(0.003)	(0.000)	(0.000)	(0.000)	(0.000)
Age at interview	0.023*	0.043***	0.12***	0.13***	-0.013	-0.0099	0.12***	0.11***
	(0.030)	(0.000)	(0.000)	(0.000)	(0.136)	(0.307)	(0.000)	(0.000)
Age at interview, squared	-0.00014	-0.00015	-0.0010***	-0.00098***	0.000088	0.000078	-0.00100***	-0.00077***
	(0.068)	(0.062)	(0.000)	(0.000)	(0.166)	(0.265)	(0.000)	(0.000)
sh_country==[2]BEL	-0.53***	-0.45***	-0.78***	-0.65***	-0.65***	-0.65***	-0.49***	-0.29***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[3]CHE	0.094**	-0.12***	0.26***	0.029	-0.36***	-0.39***	-0.011	-0.20***
	(0.005)	(0.000)	(0.000)	(0.375)	(0.000)	(0.000)	(0.758)	(0.000)
sh_country==[4]CZE	-0.99***	-0.62***	-1.44***	-1.00***	-0.42***	-0.37***	-0.24***	0.13***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[5]DEU	-0.56***	-0.46***	-0.27***	-0.15***	-0.43***	-0.42***	-0.26***	-0.14**
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.002)
sh_country==[6]DNK	0.27***	0.021	0.24***	-0.033	0.061	-0.00070	0.15***	-0.031
	(0.000)	(0.589)	(0.000)	(0.343)	(0.052)	(0.983)	(0.000)	(0.437)
sh_country==[7]ESP	-0.76***	-0.37***	-1.10***	-0.54***	-0.31***	-0.25***	-0.69***	-0.21***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[8]EST	-1.58***	-1.24***	-1.23***	-0.80***	-0.45***	-0.38***	-0.95***	-0.44***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[9]FRA	-1.01***	-0.84***	-0.54***	-0.29***	-0.58***	-0.56***	-0.68***	-0.40***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[10]HUN	-1.60***	-1.03***	-1.38***	-0.68***	-0.17***	-0.10*	-0.95***	-0.35***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.012)	(0.000)	(0.000)
sh_country==[11]ITA	-0.74***	-0.51***	-1.68***	-1.37***	-0.39***	-0.36***	-0.62***	-0.37***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[12]NLD	-0.29***	-0.35***	0.24***	0.19***	-0.62***	-0.65***	0.076	0.069
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.054)	(0.069)
sh_country==[13]POL	-0.93***	-0.36***	-1.20***	-0.53***	-0.25***	-0.17***	-1.03***	-0.45***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
sh_country==[14]PRT	-1.31***	-0.64***	-2.16***	-1.32***	-0.095*	0.098*	-1.17***	-0.35***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.011)	(0.020)	(0.000)	(0.000)
sh_country==[15]SVN	-0.89***	-0.60***	-0.17***	0.22***	-0.44***	-0.39***	-0.38***	-0.12**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.004)
sh_country==[16]SWE	0.031	-0.10*	-0.20***	-0.31***	-0.086*	-0.13***	0.024	-0.017
	(0.456)	(0.015)	(0.000)	(0.000)	(0.017)	(0.000)	(0.574)	(0.682)
Divorced/living separated		-0.10*		-0.050		-0.0082		-0.13**
		(0.034)		(0.225)		(0.849)		(0.003)
Widowed		0.067		0.10**		0.13**		-0.15***
		(0.151)		(0.010)		(0.002)		(0.001)
[1] Suburbs of big city		0.0080		0.017		0.033		-0.079**
		(0.793)		(0.528)		(0.219)		(0.010)
[2] Large town		0.038		0.019		0.077**		-0.067*
		(0.183)		(0.456)		(0.002)		(0.016)
[3] Small town		0.11***		0.066**		0.092***		0.024
		(0.000)		(0.005)		(0.000)		(0.340)
[4] Rural area/village		0.059*		0.045*		0.034		0.0038
		(0.022)		(0.048)		(0.116)		(0.876)
Employment, current job		0.18***		0.17***		0.026		0.093***
		(0.000)		(0.000)		(0.171)		(0.000)
Self-employment, current job		0.15***		0.16***		-0.018		0.066*
		(0.000)		(0.000)		(0.543)		(0.042)
[1] Primary school		0.10		0.37***		0.013		0.24***
		(0.081)		(0.000)		(0.793)		(0.000)
[2] Lower secondary school		0.14*		0.46***		0.011		0.32***
		(0.019)		(0.000)		(0.828)		(0.000)
[3] Upper secondary school		0.18**		0.58***		0.029		0.42***
		(0.002)		(0.000)		(0.568)		(0.000)
[4] Post-secondary non-tertiary education		0.25***		0.69***		0.053		0.49***
		(0.000)		(0.000)		(0.359)		(0.000)
[5] First stage tertiary education		0.26***		0.64***		0.0021		0.42***
		(0.000)		(0.000)		(0.967)		(0.000)
[6] Second stage tertiary education		0.43***		0.77***		0.016		0.42***
		(0.000)		(0.000)		(0.856)		(0.000)
[1] Fair health		1.03***		1.12***		0.15***		1.24***
		(0.000)		(0.000)		(0.000)		(0.000)
[2] Good health		1.52***		1.80***		0.17***		1.96***
		(0.000)		(0.000)		(0.000)		(0.000)
[3] Very good health		1.86***		2.18***		0.31***		2.28***
		(0.000)		(0.000)		(0.000)		(0.000)
[4] Excellent health		2.18***		2.50***		0.45***		2.40***
		(0.000)		(0.000)		(0.000)		(0.000)
Drugs for depression		-0.49***		-0.61***		-0.092***		-1.20***
		(0.000)		(0.000)		(0.000)		(0.000)
[1] Middle income		0.16***		0.19***		0.018		0.082**
		(0.000)		(0.000)		(0.429)		(0.002)
[2] Upper middle income		0.23***		0.21***		0.028		0.067**
		(0.000)		(0.000)		(0.206)		(0.008)
[3] High income		0.24***		0.25***		0.018		0.059*
		(0.000)		(0.000)		(0.402)		(0.013)
_cons	6.95***	3.91***	4.41***	0.86*	9.18***	8.73***	5.10***	2.64***
	(0.000)	(0.000)	(0.000)	(0.011)	(0.000)	(0.000)	(0.000)	(0.000)
N	52248	46969	50512	45539	52513	47161	51941	46690
R^2	0.12	0.24	0.19	0.37	0.03	0.04	0.10	0.31
adjusted R^2	0.12	0.24	0.19	0.37	0.03	0.04	0.10	0.31

Table A5.5: Regressing well-being and mental health on family status for all countries, male respondents.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
Married/registered partnership	0.57***	0.49***	0.28***	0.18**	0.40***	0.36***	0.30***	0.065
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.239)
[1] Having 1 child	0.069	-0.010	0.14**	0.035	0.16***	0.12*	0.035	0.016
	(0.167)	(0.840)	(0.004)	(0.455)	(0.000)	(0.013)	(0.469)	(0.729)
[2] Having 2 children	0.24***	0.097*	0.29***	0.13**	0.16***	0.10*	0.16***	0.094*
	(0.000)	(0.040)	(0.000)	(0.004)	(0.000)	(0.021)	(0.000)	(0.032)
[3] Having 3 or more children	0.18***	0.029	0.24***	0.056	0.13**	0.067	0.059	-0.010
	(0.000)	(0.573)	(0.000)	(0.249)	(0.004)	(0.179)	(0.235)	(0.835)
Number of resident children	-0.0084	-0.027	-0.094***	-0.11***	-0.023	-0.027	-0.018	-0.032
	(0.648)	(0.134)	(0.000)	(0.000)	(0.166)	(0.107)	(0.309)	(0.059)
Number of grandchildren	-0.0030	0.015**	-0.025***	-0.0018	0.015***	0.018***	-0.015**	-0.000094
	(0.566)	(0.004)	(0.000)	(0.707)	(0.001)	(0.000)	(0.004)	(0.985)
Controls								
Age at interview	0.055***	0.085***	0.14***	0.15***	-0.031*	-0.029	0.13***	0.12***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.024)	(0.066)	(0.000)	(0.000)
Age at interview, squared	-0.00038**	-0.00046***	-0.0012***	-0.0011***	0.00023*	0.00024*	-0.0011***	-0.00088***
	(0.002)	(0.000)	(0.000)	(0.000)	(0.023)	(0.036)	(0.000)	(0.000)
sh_country==[2]BEL	-0.46***	-0.45***	-0.73***	-0.68***	-0.59***	-0.61***	-0.34***	-0.26***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[3]CHE	0.12*	-0.14**	0.23***	-0.041	-0.29***	-0.39***	0.036	-0.19***
	(0.013)	(0.007)	(0.000)	(0.399)	(0.000)	(0.000)	(0.476)	(0.000)
sh_country==[4]CZE	-0.93***	-0.55***	-1.42***	-0.97***	-0.34***	-0.28***	-0.12*	0.23***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.017)	(0.000)
sh_country==[5]DEU	-0.55***	-0.46***	-0.30***	-0.20**	-0.41***	-0.41***	-0.13	-0.040
	(0.000)	(0.000)	(0.000)	(0.003)	(0.000)	(0.000)	(0.053)	(0.547)
sh_country==[6]DNK	0.21***	-0.062	0.18**	-0.12*	0.12*	0.020	0.19***	0.0060
	(0.000)	(0.263)	(0.001)	(0.015)	(0.015)	(0.704)	(0.001)	(0.910)
sh_country==[7]ESP	-0.62***	-0.32***	-0.89***	-0.42***	-0.29***	-0.21***	-0.23***	0.053
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.378)
sh_country==[8]EST	-1.68***	-1.35***	-1.36***	-0.91***	-0.47***	-0.43***	-0.89***	-0.44***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[9]FRA	-0.96***	-0.82***	-0.46***	-0.28***	-0.56***	-0.56***	-0.51***	-0.32***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[10]HUN	-1.55***	-1.02***	-1.32***	-0.65***	-0.16**	-0.10	-0.71***	-0.21**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.004)	(0.115)	(0.000)	(0.001)
sh_country==[11]ITA	-0.64***	-0.51***	-1.52***	-1.30***	-0.37***	-0.35***	-0.40***	-0.29***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[12]NLD	-0.24***	-0.36***	0.26***	0.15**	-0.54***	-0.59***	0.23***	0.13*
	(0.000)	(0.000)	(0.000)	(0.003)	(0.000)	(0.000)	(0.000)	(0.011)
sh_country==[13]POL	-0.85***	-0.34***	-1.13***	-0.48***	-0.23**	-0.15	-0.84***	-0.33***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.051)	(0.000)	(0.000)
sh_country==[14]PRT	-1.06***	-0.58***	-2.02***	-1.29***	-0.063	0.14*	-0.76***	-0.23**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.286)	(0.039)	(0.000)	(0.002)
sh_country==[15]SVN	-0.96***	-0.66***	-0.13*	0.25***	-0.40***	-0.37***	-0.31***	-0.072
	(0.000)	(0.000)	(0.035)	(0.000)	(0.000)	(0.000)	(0.000)	(0.224)
sh_country==[16]SWE	0.026	-0.18**	-0.29***	-0.45***	-0.050	-0.12	0.030	-0.064
	(0.670)	(0.003)	(0.000)	(0.000)	(0.392)	(0.055)	(0.623)	(0.273)
Divorced/living separated		0.0044		0.023		-0.061		-0.12
		(0.951)		(0.715)		(0.380)		(0.063)
Widowed		0.085		0.10		0.0013		-0.23**
		(0.281)		(0.143)		(0.987)		(0.001)
[1] Suburbs of big city		0.0081		0.022		0.071		-0.060
		(0.858)		(0.595)		(0.092)		(0.152)
[2] Large town		0.042		0.057		0.12**		-0.065
		(0.318)		(0.132)		(0.002)		(0.094)
[3] Small town		0.098*		0.099**		0.14***		0.049
		(0.013)		(0.005)		(0.000)		(0.165)
[4] Rural area/village		0.060		0.085*		0.13***		0.034
		(0.113)		(0.011)		(0.000)		(0.315)
Employment, current job		0.29***		0.21***		0.047		0.13***
		(0.000)		(0.000)		(0.125)		(0.000)
Self-employment, current job		0.21***		0.22***		0.023		0.074
		(0.000)		(0.000)		(0.574)		(0.059)
[1] Primary school		0.23*		0.40***		0.077		0.18*
		(0.011)		(0.000)		(0.350)		(0.034)
[2] Lower secondary school		0.24*		0.50***		0.10		0.26**
		(0.010)		(0.000)		(0.233)		(0.003)
[3] Upper secondary school		0.27**		0.62***		0.11		0.29**
		(0.003)		(0.000)		(0.177)		(0.001)
[4] Post-secondary non-tertiary education		0.37***		0.66***		0.17		0.28**
		(0.000)		(0.000)		(0.080)		(0.005)
[5] First stage tertiary education		0.32***		0.67***		0.11		0.26**
		(0.000)		(0.000)		(0.210)		(0.003)
[6] Second stage tertiary education		0.45***		0.77***		0.013		0.22
		(0.000)		(0.000)		(0.922)		(0.079)
[1] Fair health		1.05***		1.12***		0.13**		1.20***
		(0.000)		(0.000)		(0.002)		(0.000)
[2] Good health		1.49***		1.76***		0.15***		1.84***
		(0.000)		(0.000)		(0.000)		(0.000)
[3] Very good health		1.83***		2.14***		0.30***		2.12***
		(0.000)		(0.000)		(0.000)		(0.000)
[4] Excellent health		2.11***		2.45***		0.43***		2.22***
		(0.000)		(0.000)		(0.000)		(0.000)
Drugs for depression		-0.44***		-0.57***		-0.081*		-1.22***
		(0.000)		(0.000)		(0.039)		(0.000)
[1] Middle income		0.23***		0.26***		0.061		0.13***
		(0.000)		(0.000)		(0.096)		(0.000)
[2] Upper middle income		0.29***		0.31***		0.041		0.16***
		(0.000)		(0.000)		(0.253)		(0.000)
[3] High income		0.26***		0.31***		0.076*		0.11***
		(0.000)		(0.000)		(0.023)		(0.001)
Constant	5.79***	2.29***	3.65***	0.081	9.66***	9.11***	4.40***	2.41***
	(0.000)	(0.000)	(0.000)	(0.876)	(0.000)	(0.000)	(0.000)	(0.000)
N	22847	20648	22149	20067	22962	20735	22705	20518
R^2	0.12	0.25	0.18	0.36	0.037	0.046	0.061	0.27
adjusted R^2	0.12	0.25	0.18	0.36	0.036	0.044	0.06	0.27

Table A5.6: Regressing well-being and mental health on family status for all countries, female respondents.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
Married/registered partnership	0.55***	0.39***	0.30***	0.21***	0.11***	0.18***	0.27***	0.034
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.531)
[1] Having 1 child	-0.052	-0.030	0.0058	0.027	0.24***	0.20***	-0.11*	-0.030
	(0.243)	(0.521)	(0.892)	(0.512)	(0.000)	(0.000)	(0.020)	(0.518)
[2] Having 2 children	0.12**	0.13**	0.11**	0.091*	0.29***	0.25***	0.059	0.11*
	(0.004)	(0.005)	(0.008)	(0.023)	(0.000)	(0.000)	(0.195)	(0.018)
[3] Having 3 or more children	0.062	0.082	0.044	0.049	0.29***	0.25***	-0.013	0.058
	(0.201)	(0.102)	(0.334)	(0.276)	(0.000)	(0.000)	(0.807)	(0.253)
Number of resident children	-0.056**	-0.081***	-0.10***	-0.13***	-0.021	-0.027	-0.021	-0.028
	(0.004)	(0.000)	(0.000)	(0.000)	(0.150)	(0.083)	(0.310)	(0.151)
Number of grandchildren	-0.0051	0.0077	-0.019***	0.0012	0.013***	0.016***	-0.023***	-0.0040
	(0.311)	(0.117)	(0.000)	(0.774)	(0.000)	(0.000)	(0.000)	(0.400)
Controls								
Age at interview	-0.00017	0.013	0.10***	0.12***	-0.0033	-0.0031	0.11***	0.099***
	(0.991)	(0.407)	(0.000)	(0.000)	(0.760)	(0.804)	(0.000)	(0.000)
Age at interview. squared	0.000024	0.000065	-0.00095***	-0.00089***	0.0000030	0.0000075	-0.00091***	-0.00067***
	(0.819)	(0.553)	(0.000)	(0.000)	(0.971)	(0.934)	(0.000)	(0.000)
sh_country==[2]BEL	-0.60***	-0.44***	-0.82***	-0.63***	-0.70***	-0.66***	-0.60***	-0.31***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[3]CHE	0.072	-0.11*	0.29***	0.092*	-0.41***	-0.39***	-0.034	-0.20***
	(0.110)	(0.019)	(0.000)	(0.037)	(0.000)	(0.000)	(0.496)	(0.000)
sh_country==[4]CZE	-1.04***	-0.68***	-1.45***	-1.02***	-0.47***	-0.43***	-0.32***	0.061
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.215)
sh_country==[5]DEU	-0.55***	-0.46***	-0.24***	-0.11	-0.44***	-0.42***	-0.35***	-0.22***
	(0.000)	(0.000)	(0.000)	(0.060)	(0.000)	(0.000)	(0.000)	(0.001)
sh_country==[6]DNK	0.32***	0.093	0.30***	0.039	0.027	-0.0022	0.14*	-0.054
	(0.000)	(0.082)	(0.000)	(0.414)	(0.501)	(0.959)	(0.017)	(0.350)
sh_country==[7]ESP	-0.87***	-0.42***	-1.27***	-0.65***	-0.31***	-0.26***	-1.06***	-0.42***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[8]EST	-1.52***	-1.15***	-1.14***	-0.71***	-0.43***	-0.34***	-1.00***	-0.45***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[9]FRA	-1.06***	-0.86***	-0.60***	-0.30***	-0.60***	-0.55***	-0.81***	-0.46***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[10]HUN	-1.64***	-1.03***	-1.43***	-0.69***	-0.17***	-0.086	-1.13***	-0.43***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.092)	(0.000)	(0.000)
sh_country==[11]ITA	-0.82***	-0.52***	-1.81***	-1.45***	-0.40***	-0.36***	-0.79***	-0.43***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[12]NLD	-0.33***	-0.34***	0.22***	0.23***	-0.68***	-0.69***	-0.039	0.031
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.490)	(0.569)
sh_country==[13]POL	-0.99***	-0.38***	-1.25***	-0.57***	-0.25***	-0.17**	-1.17***	-0.54***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.008)	(0.000)	(0.000)
sh_country==[14]PRT	-1.51***	-0.70***	-2.28***	-1.34***	-0.12*	0.076	-1.50***	-0.45***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.016)	(0.161)	(0.000)	(0.000)
sh_country==[15]SVN	-0.85***	-0.54***	-0.20***	0.20***	-0.47***	-0.40***	-0.44***	-0.15*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.013)
sh_country==[16]SWE	0.040	-0.035	-0.13*	-0.20***	-0.10*	-0.13**	0.036	0.026
	(0.495)	(0.542)	(0.017)	(0.000)	(0.025)	(0.007)	(0.556)	(0.655)
Divorced/living separated		-0.18**		-0.10		0.0076		-0.13*
		(0.004)		(0.057)		(0.890)		(0.033)
Widowed		0.015		0.099		0.14*		-0.13*
		(0.796)		(0.058)		(0.010)		(0.029)
[1] Suburbs of big city		0.0035		0.011		0.0056		-0.098*
		(0.933)		(0.774)		(0.868)		(0.022)
[2] Large town		0.032		-0.012		0.039		-0.070
		(0.399)		(0.727)		(0.214)		(0.069)
[3] Small town		0.11**		0.036		0.053		0.00086
		(0.003)		(0.256)		(0.064)		(0.981)
[4] Rural area/village		0.056		0.0087		-0.034		-0.021
		(0.108)		(0.777)		(0.225)		(0.544)
Employment. current job		0.078*		0.13***		0.00029		0.055
		(0.013)		(0.000)		(0.991)		(0.083)
Self-employment. current job		0.084		0.087		-0.086		0.070
		(0.113)		(0.074)		(0.065)		(0.215)
[1] Primary school		0.019		0.34***		-0.018		0.25***
		(0.802)		(0.000)		(0.773)		(0.001)
[2] Lower secondary school		0.060		0.42***		-0.041		0.33***
		(0.433)		(0.000)		(0.518)		(0.000)
[3] Upper secondary school		0.11		0.53***		-0.020		0.48***
		(0.158)		(0.000)		(0.752)		(0.000)
[4] Post-secondary non-tertiary education		0.15		0.69***		-0.016		0.62***
		(0.083)		(0.000)		(0.829)		(0.000)
[5] First stage tertiary education		0.23**		0.60***		-0.073		0.51***
		(0.004)		(0.000)		(0.267)		(0.000)
[6] Second stage tertiary education		0.45**		0.76***		0.097		0.61***
		(0.001)		(0.000)		(0.384)		(0.000)
[1] Fair health		1.01***		1.12***		0.17***		1.26***
		(0.000)		(0.000)		(0.000)		(0.000)
[2] Good health		1.54***		1.82***		0.19***		2.03***
		(0.000)		(0.000)		(0.000)		(0.000)
[3] Very good health		1.88***		2.21***		0.33***		2.41***
		(0.000)		(0.000)		(0.000)		(0.000)
[4] Excellent health		2.23***		2.53***		0.47***		2.55***
		(0.000)		(0.000)		(0.000)		(0.000)
Drugs for depression		-0.51***		-0.62***		-0.10***		-1.17***
		(0.000)		(0.000)		(0.000)		(0.000)
[1] Middle income		0.10**		0.14***		-0.0070		0.054
		(0.004)		(0.000)		(0.816)		(0.145)
[2] Upper middle income		0.19***		0.13***		0.024		0.0055
		(0.000)		(0.000)		(0.409)		(0.877)
[3] High income		0.23***		0.21***		-0.015		0.037
		(0.000)		(0.000)		(0.583)		(0.271)
Constant	7.81***	5.16***	4.88***	1.40**	9.12***	8.85***	5.02***	2.37***
	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)
N	29401	26321	28363	25472	29551	26426	29236	26172
R^2	0.12	0.24	0.2	0.38	0.031	0.04	0.075	0.29
adjusted R^2	0.12	0.24	0.2	0.38	0.03	0.039	0.074	0.29

Table A5.7: Regressing well-being and mental health on family status for each country.

Country	Married				1 child				2 children				3 or more children				Resident children				Grandchildren			
	(I)	(II)	(III)	(IV)	(I)	(II)	(III)	(IV)	(I)	(II)	(III)	(IV)	(I)	(II)	(III)	(IV)	(I)	(II)	(III)	(IV)	(I)	(II)	(III)	(IV)
Austria	+	+	+	+	(+)	(+)	+	o	+	(+)	+	(+)	o	(+)	+	o	o	-	o	o	+	(-)	(+)	o
Belgium	+	+	+	+	o	o	(+)	o	o	o	+	o	o	o	+	o	o	o	o	o	o	o	o	o
Czech Republic	+	+	+	+	o	o	o	o	(+)	o	o	o	o	o	o	o	o	o	o	o	o	o	+	o
Denmark	+	+	+	o	o	o	o	o	+	o	o	o	(+)	o	o	o	o	o	o	o	o	o	o	o
Estonia	+	+	+	+	o	(+)	+	o	+	+	+	+	+	+	+	+	-	-	o	-	o	o	+	-
France	+	+	o	+	-	-	+	-	o	o	+	o	o	o	+	o	o	-	o	o	(-)	-	o	o
Germany	+	+	+	o	o	o	o	o	o	o	o	o	o	o	o	o	-	o	o	o	-	o	o	-
Hungary	+	+	+	+	o	o	+	-	o	+	+	o	o	o	+	(-)	o	-	o	o	o	-	o	-
Italy	+	+	+	+	o	o	+	-	o	o	o	-	o	(-)	o	-	o	-	o	o	o	-	o	-
Netherlands	+	+	o	+	o	o	(+)	o	+	o	(+)	o	+	o	o	o	o	o	o	o	o	o	(+)	o
Poland	+	(+)	+	o	o	o	o	o	o	o	o	o	o	o	o	o	o	-	o	o	o	o	+	o
Portugal	+	o	+	o	o	o	o	o	(+)	(+)	o	o	(+)	+	o	o	o	-	o	o	-	-	o	-
Slovenia	+	+	(+)	+	o	o	+	o	o	o	o	o	o	o	o	o	o	o	o	o	o	(-)	+	(-)
Spain	+	+	+	+	o	o	+	o	o	+	+	o	o	o	+	o	o	-	o	-	o	-	o	-
Sweden	+	+	+	o	o	+	+	o	o	+	(+)	+	o	+	+	o	o	o	(+)	o	o	o	o	o
Switzerland	+	+	o	+	o	o	o	o	o	o	+	(+)	+	+	+	o	o	o	o	o	o	o	o	o

Notes: +/ - indicates positive/negative significant effect at 5% significance level; additionally (+),(-) indicates positive, or negative effect at 10% significance level, and o indicates that there is no significant effect at 10% level. Dependent Variables: (I) Life satisfaction, (II) CASP-12, (III) Network satisfaction, (IV) EURO-D. (I)-(IV) country-specific OLS Regression. Controls A: female, age, age2, country dummy. Children: A dummy variable for having no children (excluded category), one child, two children, and three or more children. Resident children: Number of children living with their parents. If a respondent has no children then the value is set to 0. Grandchildren: Number of grandchildren, Married: Dummy variable if respondent is married or in registered partnership. Excluded category: Married but living separated from a spouse, never married, divorced, widowed.

Table A5.8: Regressing well-being and mental health on network types controlling for network size and family status for all countries, all respondents.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
[1] Partner	0.44***	0.46***	0.20***	0.20***	2.37***	2.47***	0.30***	0.32***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
[2] Children	0.21***	0.28***	-0.082	0.028	2.42***	2.50***	0.021	0.19***
	(0.001)	(0.000)	(0.123)	(0.572)	(0.000)	(0.000)	(0.723)	(0.001)
[3] Other Relatives	0.21***	0.26***	0.066	0.12*	2.14***	2.24***	0.0040	0.098
	(0.000)	(0.000)	(0.219)	(0.014)	(0.000)	(0.000)	(0.946)	(0.086)
[4] Family	0.30***	0.37***	0.037	0.12*	2.30***	2.40***	0.12*	0.25***
	(0.000)	(0.000)	(0.477)	(0.014)	(0.000)	(0.000)	(0.034)	(0.000)
[5] Friends	0.21***	0.24***	0.10	0.15**	2.03***	2.13***	0.0013	0.095
	(0.000)	(0.000)	(0.051)	(0.002)	(0.000)	(0.000)	(0.982)	(0.088)
[6] Diverse	0.12	0.23***	-0.083	0.045	1.97***	2.07***	-0.12*	0.045
	(0.055)	(0.000)	(0.135)	(0.385)	(0.000)	(0.000)	(0.044)	(0.440)
Size of social network	0.097***	0.065***	0.11***	0.069***	0.057***	0.054***	0.053***	0.017**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.004)
Married/registered partnership	0.48***	0.38***	0.24***	0.17***	0.10***	0.13***	0.20***	-0.022
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.576)
[1] Having 1 child	-0.018	-0.054	0.087**	0.035	0.10***	0.074*	-0.056	-0.047
	(0.589)	(0.125)	(0.006)	(0.260)	(0.000)	(0.014)	(0.105)	(0.172)
[2] Having 2 children	0.13***	0.062	0.19***	0.097**	0.095***	0.069*	0.075*	0.055
	(0.000)	(0.064)	(0.000)	(0.001)	(0.000)	(0.019)	(0.023)	(0.092)
[3] Having 3 or more children	0.055	0.0015	0.12***	0.036	0.077**	0.046	-0.018	-0.026
	(0.124)	(0.967)	(0.000)	(0.280)	(0.008)	(0.149)	(0.619)	(0.473)
Number of resident children	-0.027*	-0.050***	-0.094***	-0.12***	-0.015	-0.021	-0.015	-0.024
	(0.045)	(0.000)	(0.000)	(0.000)	(0.135)	(0.057)	(0.281)	(0.058)
Number of grandchildren	-0.0054	0.0093**	-0.021***	-0.00045	0.012***	0.015***	-0.021***	-0.0032
	(0.138)	(0.009)	(0.000)	(0.884)	(0.000)	(0.000)	(0.000)	(0.349)
Controls								
Female	-0.044**	0.041**	-0.17***	-0.060***	0.11***	0.11***	-0.64***	-0.50***
	(0.005)	(0.009)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age at interview	0.020	0.041***	0.11***	0.13***	-0.013	-0.011	0.12***	0.11***
	(0.064)	(0.000)	(0.000)	(0.000)	(0.099)	(0.219)	(0.000)	(0.000)
Age at interview, squared	-0.00011	-0.00013	-0.00099***	-0.00095***	0.000088	0.000086	-0.00097***	-0.00076***
	(0.161)	(0.110)	(0.000)	(0.000)	(0.140)	(0.192)	(0.000)	(0.000)
sh_country==[2]BEL	-0.53***	-0.44***	-0.78***	-0.65***	-0.63***	-0.63***	-0.49***	-0.29***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[3]CHE	0.089**	-0.12***	0.26***	0.026	-0.37***	-0.41***	-0.022	-0.22***
	(0.008)	(0.000)	(0.000)	(0.424)	(0.000)	(0.000)	(0.533)	(0.000)
sh_country==[4]CZE	-0.94***	-0.60***	-1.38***	-0.97***	-0.37***	-0.35***	-0.24***	0.12**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
sh_country==[5]DEU	-0.56***	-0.47***	-0.26***	-0.15**	-0.45***	-0.43***	-0.28***	-0.16***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[6]DNK	0.27***	0.033	0.26***	-0.015	0.035	-0.014	0.15***	-0.038
	(0.000)	(0.398)	(0.000)	(0.660)	(0.272)	(0.685)	(0.000)	(0.335)
sh_country==[7]ESP	-0.75***	-0.38***	-1.07***	-0.54***	-0.33***	-0.30***	-0.70***	-0.23***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[8]EST	-1.55***	-1.21***	-1.19***	-0.77***	-0.42***	-0.36***	-0.94***	-0.44***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[9]FRA	-0.98***	-0.82***	-0.52***	-0.29***	-0.51***	-0.50***	-0.67***	-0.38***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[10]HUN	-1.58***	-1.03***	-1.34***	-0.67***	-0.18***	-0.12**	-0.94***	-0.35***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
sh_country==[11]ITA	-0.67***	-0.47***	-1.60***	-1.34***	-0.25***	-0.22***	-0.58***	-0.35***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[12]NLD	-0.28***	-0.35***	0.25***	0.20***	-0.64***	-0.67***	0.075	0.063
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.058)	(0.095)
sh_country==[13]POL	-0.89***	-0.35***	-1.13***	-0.50***	-0.25***	-0.18***	-1.04***	-0.49***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[14]PRT	-1.30***	-0.66***	-2.13***	-1.32***	-0.13***	0.036	-1.19***	-0.38***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.364)	(0.000)	(0.000)
sh_country==[15]SVN	-0.82***	-0.55***	-0.091*	0.27***	-0.31***	-0.26***	-0.39***	-0.13**
	(0.000)	(0.000)	(0.024)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)
sh_country==[16]SWE	0.045	-0.094*	-0.17***	-0.29***	-0.14***	-0.18***	0.019	-0.034
	(0.288)	(0.027)	(0.000)	(0.000)	(0.000)	(0.000)	(0.667)	(0.416)
Divorced/living separated		-0.11*		-0.047		-0.021		-0.14**
		(0.025)		(0.250)		(0.600)		(0.002)
Widowed		0.072		0.13**		0.091*		-0.15***
		(0.123)		(0.002)		(0.024)		(0.001)

Table A5.8 (continued): Regressing well-being and mental health on network types controlling for network size and family status for all countries, all respondents.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
[1] Suburbs of big city		0.013		0.032		0.015		-0.087**
		(0.681)		(0.248)		(0.546)		(0.004)
[2] Large town		0.047		0.037		0.077**		-0.074**
		(0.098)		(0.139)		(0.001)		(0.008)
[3] Small town		0.11***		0.080***		0.089***		0.018
		(0.000)		(0.001)		(0.000)		(0.490)
[4] Rural area/village		0.069**		0.066**		0.033		-0.0040
		(0.007)		(0.003)		(0.116)		(0.871)
Employment, current job		0.17***		0.17***		0.025		0.097***
		(0.000)		(0.000)		(0.180)		(0.000)
Self-employment, current job		0.15***		0.16***		-0.021		0.072*
		(0.000)		(0.000)		(0.475)		(0.026)
[1] Primary school		0.087		0.35***		-0.034		0.23***
		(0.135)		(0.000)		(0.454)		(0.000)
[2] Lower secondary school		0.12*		0.44***		-0.044		0.32***
		(0.048)		(0.000)		(0.340)		(0.000)
[3] Upper secondary school		0.15*		0.54***		-0.041		0.41***
		(0.012)		(0.000)		(0.380)		(0.000)
[4] Post-secondary non-tertiary education		0.21**		0.64***		-0.018		0.49***
		(0.002)		(0.000)		(0.735)		(0.000)
[5] First stage tertiary education		0.22***		0.58***		-0.070		0.42***
		(0.000)		(0.000)		(0.139)		(0.000)
[6] Second stage tertiary education		0.39***		0.71***		-0.050		0.42***
		(0.000)		(0.000)		(0.545)		(0.000)
[1] Fair health		1.02***		1.11***		0.14***		1.24***
		(0.000)		(0.000)		(0.000)		(0.000)
[2] Good health		1.51***		1.78***		0.16***		1.95***
		(0.000)		(0.000)		(0.000)		(0.000)
[3] Very good health		1.84***		2.16***		0.30***		2.28***
		(0.000)		(0.000)		(0.000)		(0.000)
[4] Excellent health		2.17***		2.48***		0.44***		2.40***
		(0.000)		(0.000)		(0.000)		(0.000)
Drugs for depression		-0.50***		-0.61***		-0.094***		-1.19***
		(0.000)		(0.000)		(0.000)		(0.000)
[1] Middle income		0.15***		0.19***		-0.0028		0.080**
		(0.000)		(0.000)		(0.898)		(0.002)
[2] Upper middle income		0.23***		0.21***		0.016		0.070**
		(0.000)		(0.000)		(0.469)		(0.006)
[3] High income		0.24***		0.24***		0.00069		0.061*
		(0.000)		(0.000)		(0.973)		(0.010)
_cons	6.61***	3.62***	4.20***	0.72*	7.13***	6.70***	5.04***	2.55***
	(0.000)	(0.000)	(0.000)	(0.035)	(0.000)	(0.000)	(0.000)	(0.000)
N	52248	46969	50512	45539	52513	47161	51941	46690
R^2	0.13	0.25	0.20	0.38	0.12	0.13	0.10	0.31
adjusted R^2	0.13	0.25	0.20	0.38	0.12	0.13	0.10	0.31

Table A5.9: Regressing well-being and mental health on network types controlling for network size and family status for all countries, male respondents.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
[1] Partner	0.46***	0.44***	0.16*	0.16*	2.28***	2.39***	0.19**	0.22**
	(0.000)	(0.000)	(0.019)	(0.015)	(0.000)	(0.000)	(0.008)	(0.001)
[2] Children	0.32***	0.33***	0.038	0.11	2.27***	2.38***	0.059	0.20*
	(0.000)	(0.000)	(0.616)	(0.141)	(0.000)	(0.000)	(0.472)	(0.011)
[3] Other Relatives	0.26**	0.27***	0.098	0.14	1.94***	2.05***	0.0095	0.094
	(0.001)	(0.001)	(0.191)	(0.052)	(0.000)	(0.000)	(0.905)	(0.226)
[4] Family	0.37***	0.40***	0.054	0.12	2.15***	2.26***	0.089	0.20**
	(0.000)	(0.000)	(0.455)	(0.086)	(0.000)	(0.000)	(0.245)	(0.006)
[5] Friends	0.31***	0.30***	0.15*	0.17*	1.80***	1.92***	0.014	0.084
	(0.000)	(0.000)	(0.037)	(0.021)	(0.000)	(0.000)	(0.856)	(0.271)
[6] Diverse	0.23**	0.23**	-0.0027	0.050	1.70***	1.81***	-0.043	0.057
	(0.007)	(0.006)	(0.972)	(0.511)	(0.000)	(0.000)	(0.599)	(0.474)
Size of social network	0.076***	0.052***	0.089***	0.055***	0.050***	0.048***	0.041***	0.016
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.081)
Married/registered partnership	0.50***	0.43***	0.25***	0.17**	0.22***	0.15**	0.24***	0.019
	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)	(0.007)	(0.000)	(0.739)
[1] Having 1 child	0.040	-0.038	0.14**	0.035	0.056	0.041	0.019	-0.011
	(0.427)	(0.473)	(0.003)	(0.452)	(0.175)	(0.365)	(0.699)	(0.813)
[2] Having 2 children	0.20***	0.058	0.28***	0.12**	0.029	0.014	0.13**	0.060
	(0.000)	(0.229)	(0.000)	(0.008)	(0.452)	(0.753)	(0.004)	(0.179)
[3] Having 3 or more children	0.12*	-0.016	0.22***	0.040	-0.0073	-0.035	0.025	-0.047
	(0.019)	(0.757)	(0.000)	(0.414)	(0.866)	(0.457)	(0.617)	(0.344)
Number of resident children	-0.0067	-0.025	-0.094***	-0.11***	-0.020	-0.024	-0.017	-0.030
	(0.713)	(0.162)	(0.000)	(0.000)	(0.199)	(0.129)	(0.339)	(0.072)
Number of grandchildren	-0.0033	0.014**	-0.025***	-0.0020	0.013**	0.016***	-0.016**	-0.00068
	(0.528)	(0.007)	(0.000)	(0.672)	(0.002)	(0.000)	(0.003)	(0.889)
Controls								
Age at interview	0.052**	0.083***	0.14***	0.15***	-0.031*	-0.025	0.13***	0.12***
	(0.002)	(0.000)	(0.000)	(0.000)	(0.016)	(0.080)	(0.000)	(0.000)
Age at interview, squared	-0.00036**	-0.00044***	-0.0012***	-0.0011***	0.00022*	0.00021*	-0.0011***	-0.00087***
	(0.003)	(0.000)	(0.000)	(0.000)	(0.019)	(0.048)	(0.000)	(0.000)
sh_country==[2]BEL	-0.43***	-0.43***	-0.71***	-0.68***	-0.54***	-0.56***	-0.32***	-0.25***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[3]CHE	0.13**	-0.13*	0.23***	-0.048	-0.28***	-0.39***	0.042	-0.19***
	(0.007)	(0.010)	(0.000)	(0.325)	(0.000)	(0.000)	(0.419)	(0.000)
sh_country==[4]CZE	-0.88***	-0.53***	-1.37***	-0.95***	-0.32***	-0.27***	-0.11*	0.23***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.032)	(0.000)
sh_country==[5]DEU	-0.54***	-0.46***	-0.28***	-0.19**	-0.41***	-0.41***	-0.13	-0.047
	(0.000)	(0.000)	(0.000)	(0.003)	(0.000)	(0.000)	(0.063)	(0.478)
sh_country==[6]DNK	0.24***	-0.038	0.20***	-0.11*	0.11*	0.017	0.20***	0.0077
	(0.000)	(0.500)	(0.000)	(0.030)	(0.021)	(0.751)	(0.001)	(0.886)
sh_country==[7]ESP	-0.61***	-0.33***	-0.87***	-0.42***	-0.31***	-0.27***	-0.23***	0.045
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.459)
sh_country==[8]EST	-1.65***	-1.33***	-1.32***	-0.90***	-0.46***	-0.43***	-0.88***	-0.44***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[9]FRA	-0.91***	-0.79***	-0.44***	-0.28***	-0.46***	-0.47***	-0.49***	-0.31***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[10]HUN	-1.53***	-1.02***	-1.29***	-0.65***	-0.17**	-0.12*	-0.70***	-0.21***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.045)	(0.000)	(0.001)
sh_country==[11]ITA	-0.58***	-0.46***	-1.46***	-1.27***	-0.21***	-0.19***	-0.37***	-0.27***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[12]NLD	-0.22***	-0.35***	0.28***	0.16**	-0.55***	-0.60***	0.24***	0.13**
	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)	(0.010)
sh_country==[13]POL	-0.81***	-0.33***	-1.07***	-0.46***	-0.27***	-0.21**	-0.84***	-0.35***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.004)	(0.000)	(0.000)
sh_country==[14]PRT	-1.05***	-0.59***	-2.00***	-1.30***	-0.083	0.080	-0.76***	-0.24**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.133)	(0.208)	(0.000)	(0.001)
sh_country==[15]SVN	-0.89***	-0.63***	-0.052	0.29***	-0.30***	-0.29***	-0.30***	-0.068
	(0.000)	(0.000)	(0.393)	(0.000)	(0.000)	(0.000)	(0.000)	(0.253)
sh_country==[16]SWE	0.059	-0.16*	-0.25***	-0.43***	-0.11	-0.18**	0.041	-0.073
	(0.336)	(0.011)	(0.000)	(0.000)	(0.060)	(0.004)	(0.501)	(0.222)
Divorced/living separated		-0.0038		0.027		-0.094		-0.13*
		(0.958)		(0.663)		(0.151)		(0.049)
Widowed		0.085		0.11		-0.054		-0.25***

Table A5.9 (continued): Regressing well-being and mental health on network types controlling for network size and family status for all countries, male respondents.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
[1] Suburbs of big city		0.010		0.033		0.038		-0.066
		(0.817)		(0.420)		(0.344)		(0.114)
[2] Large town		0.049		0.072		0.11**		-0.071
		(0.252)		(0.057)		(0.003)		(0.070)
[3] Small town		0.10**		0.11**		0.13***		0.044
		(0.010)		(0.002)		(0.000)		(0.210)
[4] Rural area/village		0.068		0.10**		0.12***		0.029
		(0.073)		(0.002)		(0.000)		(0.398)
Employment, current job		0.29***		0.22***		0.058*		0.13***
		(0.000)		(0.000)		(0.044)		(0.000)
Self-employment, current job		0.21***		0.21***		0.032		0.078*
		(0.000)		(0.000)		(0.407)		(0.045)
[1] Primary school		0.21*		0.39***		0.019		0.18*
		(0.019)		(0.000)		(0.798)		(0.038)
[2] Lower secondary school		0.22*		0.48***		0.037		0.26**
		(0.019)		(0.000)		(0.634)		(0.003)
[3] Upper secondary school		0.24**		0.59***		0.037		0.28**
		(0.007)		(0.000)		(0.628)		(0.001)
[4] Post-secondary non-tertiary education		0.34**		0.62***		0.070		0.27**
		(0.001)		(0.000)		(0.421)		(0.005)
[5] First stage tertiary education		0.29**		0.63***		0.034		0.26**
		(0.002)		(0.000)		(0.659)		(0.003)
[6] Second stage tertiary education		0.41**		0.72***		-0.064		0.23
		(0.002)		(0.000)		(0.605)		(0.080)
[1] Fair health		1.05***		1.12***		0.11**		1.20***
		(0.000)		(0.000)		(0.004)		(0.000)
[2] Good health		1.48***		1.75***		0.14***		1.84***
		(0.000)		(0.000)		(0.000)		(0.000)
[3] Very good health		1.82***		2.13***		0.28***		2.12***
		(0.000)		(0.000)		(0.000)		(0.000)
[4] Excellent health		2.10***		2.44***		0.41***		2.22***
		(0.000)		(0.000)		(0.000)		(0.000)
Drugs for depression		-0.44***		-0.58***		-0.078*		-1.22***
		(0.000)		(0.000)		(0.036)		(0.000)
[1] Middle income		0.22***		0.26***		0.037		0.13***
		(0.000)		(0.000)		(0.289)		(0.000)
[2] Upper middle income		0.29***		0.30***		0.022		0.16***
		(0.000)		(0.000)		(0.522)		(0.000)
[3] High income		0.26***		0.30***		0.057		0.11***
		(0.000)		(0.000)		(0.071)		(0.001)
_cons	5.42***	1.99**	3.37***	-0.14	7.81***	7.14***	4.35***	2.33***
	(0.000)	(0.001)	(0.000)	(0.786)	(0.000)	(0.000)	(0.000)	(0.000)
N	22847	20648	22149	20067	22962	20735	22705	20518
R^2	0.13	0.25	0.18	0.36	0.13	0.14	0.06	0.27
adjusted R^2	0.13	0.25	0.18	0.36	0.13	0.14	0.06	0.27

Table A5.10: Regressing well-being and mental health on network types controlling for network size and family status for all countries, female respondents.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
[1] Partner	0.45***	0.50***	0.29***	0.29***	2.38***	2.48***	0.48***	0.46***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
[2] Children	0.15	0.27**	-0.11	0.022	2.57***	2.65***	0.041	0.22**
	(0.087)	(0.002)	(0.136)	(0.747)	(0.000)	(0.000)	(0.641)	(0.007)
[3] Other Relatives	0.16	0.26**	0.061	0.13	2.34***	2.43***	0.020	0.12
	(0.058)	(0.003)	(0.426)	(0.061)	(0.000)	(0.000)	(0.819)	(0.135)
[4] Family	0.25**	0.36***	0.044	0.14*	2.47***	2.55***	0.16	0.30***
	(0.003)	(0.000)	(0.559)	(0.050)	(0.000)	(0.000)	(0.060)	(0.000)
[5] Friends	0.13	0.21*	0.082	0.17*	2.25***	2.34***	0.0080	0.12
	(0.133)	(0.013)	(0.279)	(0.016)	(0.000)	(0.000)	(0.928)	(0.128)
[6] Diverse	0.041	0.23**	-0.11	0.066	2.20***	2.29***	-0.15	0.063
	(0.647)	(0.009)	(0.157)	(0.363)	(0.000)	(0.000)	(0.095)	(0.458)
Size of social network	0.11***	0.074***	0.12***	0.077***	0.058***	0.056***	0.059***	0.016*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.040)
Married/registered partnership	0.47***	0.33***	0.22***	0.17***	0.054**	0.12*	0.16***	-0.034
	(0.000)	(0.000)	(0.000)	(0.001)	(0.004)	(0.011)	(0.000)	(0.536)
[1] Having 1 child	-0.070	-0.060	0.039	0.040	0.11**	0.089*	-0.12*	-0.062
	(0.124)	(0.203)	(0.372)	(0.348)	(0.004)	(0.029)	(0.015)	(0.195)
[2] Having 2 children	0.068	0.070	0.11**	0.085*	0.12***	0.10**	0.027	0.065
	(0.119)	(0.131)	(0.008)	(0.040)	(0.001)	(0.010)	(0.559)	(0.163)
[3] Having 3 or more children	-0.0012	0.017	0.042	0.036	0.13**	0.11*	-0.055	0.0093
	(0.980)	(0.742)	(0.368)	(0.435)	(0.001)	(0.015)	(0.306)	(0.859)
Number of resident children	-0.048*	-0.075***	-0.095***	-0.13***	-0.020	-0.025	-0.010	-0.021
	(0.012)	(0.000)	(0.000)	(0.000)	(0.148)	(0.084)	(0.616)	(0.276)
Number of grandchildren	-0.0056	0.0062	-0.018***	0.00098	0.010**	0.012**	-0.024***	-0.0052
	(0.264)	(0.203)	(0.000)	(0.812)	(0.005)	(0.001)	(0.000)	(0.276)
Controls								
Age at interview	-0.0037	0.0096	0.097***	0.11***	-0.0057	-0.0085	0.11***	0.100***
	(0.792)	(0.526)	(0.000)	(0.000)	(0.586)	(0.467)	(0.000)	(0.000)
Age at interview, squared	0.000064	0.000092	-0.00088***	-0.00084***	0.000024	0.000050	-0.00089***	-0.00068***
	(0.535)	(0.401)	(0.000)	(0.000)	(0.758)	(0.554)	(0.000)	(0.000)
sh_country==[2]BEL	-0.61***	-0.45***	-0.84***	-0.63***	-0.70***	-0.67***	-0.62***	-0.31***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[3]CHE	0.045	-0.12*	0.27***	0.088	-0.44***	-0.42***	-0.071	-0.23***
	(0.326)	(0.011)	(0.000)	(0.050)	(0.000)	(0.000)	(0.157)	(0.000)
sh_country==[4]CZE	-0.99***	-0.65***	-1.39***	-0.99***	-0.41***	-0.39***	-0.34***	0.036
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.467)
sh_country==[5]DEU	-0.57***	-0.47***	-0.24***	-0.10	-0.47***	-0.44***	-0.39***	-0.25***
	(0.000)	(0.000)	(0.000)	(0.087)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[6]DNK	0.30***	0.094	0.29***	0.057	-0.023	-0.030	0.11	-0.076
	(0.000)	(0.082)	(0.000)	(0.243)	(0.581)	(0.492)	(0.067)	(0.193)
sh_country==[7]ESP	-0.86***	-0.42***	-1.24***	-0.65***	-0.35***	-0.31***	-1.08***	-0.45***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[8]EST	-1.49***	-1.13***	-1.11***	-0.68***	-0.41***	-0.30***	-1.00***	-0.46***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[9]FRA	-1.03***	-0.84***	-0.59***	-0.30***	-0.55***	-0.52***	-0.80***	-0.45***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[10]HUN	-1.62***	-1.02***	-1.37***	-0.67***	-0.20***	-0.11*	-1.11***	-0.43***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.019)	(0.000)	(0.000)
sh_country==[11]ITA	-0.74***	-0.48***	-1.71***	-1.40***	-0.26***	-0.23***	-0.74***	-0.40***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[12]NLD	-0.34***	-0.35***	0.22***	0.23***	-0.71***	-0.72***	-0.059	0.015
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.303)	(0.777)
sh_country==[13]POL	-0.95***	-0.37***	-1.18***	-0.53***	-0.23***	-0.15*	-1.21***	-0.58***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.015)	(0.000)	(0.000)
sh_country==[14]PRT	-1.51***	-0.72***	-2.25***	-1.34***	-0.16***	0.012	-1.54***	-0.50***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.819)	(0.000)	(0.000)
sh_country==[15]SVN	-0.77***	-0.49***	-0.13*	0.25***	-0.30***	-0.22***	-0.46***	-0.17**
	(0.000)	(0.000)	(0.019)	(0.000)	(0.000)	(0.000)	(0.000)	(0.004)
sh_country==[16]SWE	0.024	-0.038	-0.12*	-0.18***	-0.16***	-0.18***	0.0013	-0.0066
	(0.678)	(0.511)	(0.025)	(0.001)	(0.001)	(0.000)	(0.983)	(0.911)
Divorced/living separated		-0.18**		-0.10		0.014		-0.13*
		(0.003)		(0.058)		(0.786)		(0.032)
Widowed		0.023		0.12*		0.12*		-0.12*
		(0.705)		(0.025)		(0.020)		(0.037)

Table A5.10 (continued): Regressing well-being and mental health on network types controlling for network size and family status for all countries, female respondents.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
[1] Suburbs of big city		0.010		0.028		0.00070		-0.11*
		(0.802)		(0.443)		(0.983)		(0.013)
[2] Large town		0.043		0.0086		0.048		-0.078*
		(0.256)		(0.798)		(0.109)		(0.042)
[3] Small town		0.11**		0.052		0.057*		-0.0069
		(0.002)		(0.097)		(0.040)		(0.848)
[4] Rural area/village		0.068		0.033		-0.029		-0.032
		(0.053)		(0.275)		(0.279)		(0.359)
Employment, current job		0.075*		0.13***		-0.0089		0.063*
		(0.017)		(0.000)		(0.709)		(0.048)
Self-employment, current job		0.084		0.082		-0.098*		0.077
		(0.111)		(0.092)		(0.034)		(0.175)
[1] Primary school		0.0061		0.32***		-0.061		0.25***
		(0.935)		(0.000)		(0.275)		(0.001)
[2] Lower secondary school		0.039		0.39***		-0.097		0.33***
		(0.609)		(0.000)		(0.095)		(0.000)
[3] Upper secondary school		0.076		0.49***		-0.091		0.48***
		(0.321)		(0.000)		(0.118)		(0.000)
[4] Post-secondary non-tertiary education		0.12		0.64***		-0.078		0.62***
		(0.189)		(0.000)		(0.260)		(0.000)
[5] First stage tertiary education		0.17*		0.53***		-0.15*		0.51***
		(0.025)		(0.000)		(0.012)		(0.000)
[6] Second stage tertiary education		0.42**		0.71***		0.065		0.61***
		(0.003)		(0.000)		(0.558)		(0.000)
[1] Fair health		1.00***		1.10***		0.15***		1.27***
		(0.000)		(0.000)		(0.000)		(0.000)
[2] Good health		1.53***		1.80***		0.19***		2.03***
		(0.000)		(0.000)		(0.000)		(0.000)
[3] Very good health		1.86***		2.19***		0.31***		2.40***
		(0.000)		(0.000)		(0.000)		(0.000)
[4] Excellent health		2.22***		2.51***		0.47***		2.54***
		(0.000)		(0.000)		(0.000)		(0.000)
Drugs for depression		-0.51***		-0.63***		-0.11***		-1.16***
		(0.000)		(0.000)		(0.000)		(0.000)
[1] Middle income		0.095**		0.13***		-0.026		0.051
		(0.007)		(0.000)		(0.362)		(0.170)
[2] Upper middle income		0.18***		0.13***		0.011		0.010
		(0.000)		(0.000)		(0.690)		(0.770)
[3] High income		0.22***		0.20***		-0.036		0.041
		(0.000)		(0.000)		(0.174)		(0.225)
_cons	7.46***	4.85***	4.65***	1.25**	6.88***	6.70***	4.91***	2.21***
	(0.000)	(0.000)	(0.000)	(0.006)	(0.000)	(0.000)	(0.000)	(0.000)
N	29401	26321	28363	25472	29551	26426	29236	26172
R^2	0.13	0.25	0.21	0.38	0.12	0.13	0.081	0.3
adjusted R^2	0.13	0.25	0.21	0.38	0.12	0.13	0.08	0.29

Table A5.11: Well-being and mental health measures conditional on network size over all countries.

	Life satisfaction			Quality of life (CASP-12)			Network satisfaction			Lack of depressive symptoms (EURO-D)		
Size	All (1)	Male (2)	Female (3)	All (1)	Male (2)	Female (3)	All (1)	Male (2)	Female (3)	All (1)	Male (2)	Female (3)
0	6.93	6.98	6.88	6.48	6.67	6.29	6.49	6.60	6.38	7.59	7.99	7.18
1	7.45	7.58	7.32	6.80	6.97	6.61	8.86	8.89	8.83	7.89	8.25	7.49
2	7.48	7.61	7.38	6.84	7.03	6.69	8.90	8.83	8.95	7.81	8.26	7.46
3	7.62	7.75	7.54	7.07	7.21	6.97	8.95	8.86	9.02	7.83	8.26	7.56
4	7.76	7.83	7.72	7.18	7.27	7.12	8.95	8.82	9.03	7.90	8.34	7.63
5	7.8	7.92	7.76	7.34	7.50	7.26	8.96	8.87	9.00	7.92	8.43	7.66
6	7.99	8.06	7.96	7.45	7.55	7.39	8.96	8.82	9.03	8.01	8.46	7.80
7	8.04	8.06	8.03	7.46	7.44	7.47	8.99	8.84	9.07	8.01	8.35	7.83

Notes: For each well-being and mental health measure column (1)-(3) represent the average conditional on the network size for all respondents and by gender.

Table A5.12: Regressing well-being and mental health on network types controlling for network size, relational dynamics and family status for all countries, all respondents with social support network.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
[2] Children	-0.22***	-0.12***	-0.29***	-0.15***	0.14***	0.12***	-0.28***	-0.084*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.011)
[3] Other Relatives	-0.12***	-0.059	-0.053	0.025	0.097***	0.092***	-0.25***	-0.14***
	(0.000)	(0.072)	(0.093)	(0.400)	(0.000)	(0.000)	(0.000)	(0.000)
[4] Family	-0.10***	-0.023	-0.14***	-0.039	0.076***	0.068***	-0.16***	-0.022
	(0.000)	(0.403)	(0.000)	(0.115)	(0.000)	(0.001)	(0.000)	(0.422)
[5] Friends	-0.052	-0.029	0.057	0.11***	0.13***	0.12***	-0.21***	-0.12***
	(0.101)	(0.352)	(0.058)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
[6] Diverse	-0.15***	-0.055	-0.13***	-0.011	0.043	0.032	-0.34***	-0.18***
	(0.000)	(0.109)	(0.000)	(0.713)	(0.085)	(0.220)	(0.000)	(0.000)
Size of social network	0.099***	0.074***	0.11***	0.074***	0.073***	0.072***	0.052***	0.024***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average contact 0-6	0.064***	0.066***	0.068***	0.066***	0.18***	0.17***	0.032**	0.030**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.009)	(0.008)
Average closeness 0-3	0.32***	0.26***	0.30***	0.24***	0.70***	0.68***	0.17***	0.11***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average proximity 0-5	-0.041***	-0.0045	-0.076***	-0.030***	-0.052***	-0.053***	-0.033**	0.019
	(0.000)	(0.657)	(0.000)	(0.001)	(0.000)	(0.000)	(0.002)	(0.063)
Married/registered partnership	0.44***	0.33***	0.21***	0.14***	-0.013	0.027	0.18***	-0.042
	(0.000)	(0.000)	(0.000)	(0.000)	(0.338)	(0.385)	(0.000)	(0.290)
[1] Having 1 child	-0.050	-0.083*	0.056	0.0039	0.023	-0.00071	-0.069*	-0.065
	(0.140)	(0.018)	(0.082)	(0.901)	(0.332)	(0.978)	(0.049)	(0.060)
[2] Having 2 children	0.098**	0.035	0.15***	0.064*	0.028	0.0041	0.058	0.037
	(0.002)	(0.291)	(0.000)	(0.036)	(0.216)	(0.872)	(0.083)	(0.264)
[3] Having 3 or more children	0.028	-0.019	0.086*	0.0080	0.019	-0.0071	-0.034	-0.038
	(0.438)	(0.610)	(0.013)	(0.811)	(0.439)	(0.797)	(0.368)	(0.294)
Number of resident children	-0.030*	-0.058***	-0.089***	-0.12***	-0.027**	-0.030**	-0.010	-0.030*
	(0.025)	(0.000)	(0.000)	(0.000)	(0.003)	(0.002)	(0.459)	(0.024)
Number of grandchildren	-0.0072	0.0064	-0.021***	-0.0016	0.0095***	0.010***	-0.021***	-0.0046
	(0.052)	(0.077)	(0.000)	(0.600)	(0.000)	(0.000)	(0.000)	(0.188)
Controls								
Female	-0.074***	0.016	-0.19***	-0.083***	0.055***	0.051***	-0.65***	-0.51***
	(0.000)	(0.319)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age at interview	0.028*	0.047***	0.12***	0.13***	0.0039	0.0032	0.12***	0.11***
	(0.010)	(0.000)	(0.000)	(0.000)	(0.583)	(0.686)	(0.000)	(0.000)
Age at interview, squared	-0.00016*	-0.00018*	-0.0011***	-0.00100***	-0.000013	-0.0000018	-0.0010***	-0.00079***
	(0.047)	(0.032)	(0.000)	(0.000)	(0.806)	(0.975)	(0.000)	(0.000)
sh_country==[2]BEL	-0.37***	-0.31***	-0.62***	-0.52***	-0.27***	-0.28***	-0.40***	-0.23***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[3]CHE	0.35***	0.10**	0.51***	0.25***	0.21***	0.20***	0.100**	-0.12**
	(0.000)	(0.006)	(0.000)	(0.000)	(0.000)	(0.000)	(0.009)	(0.001)
sh_country==[4]CZE	-0.80***	-0.50***	-1.25***	-0.88***	-0.044	-0.046	-0.16***	0.17***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.092)	(0.121)	(0.000)	(0.000)
sh_country==[5]DEU	-0.33***	-0.27***	-0.036	0.038	0.067	0.067	-0.15**	-0.076
	(0.000)	(0.000)	(0.464)	(0.405)	(0.076)	(0.093)	(0.003)	(0.114)
sh_country==[6]DNK	0.44***	0.18***	0.39***	0.11**	0.36***	0.32***	0.22***	0.017
	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)	(0.665)
sh_country==[7]ESP	-0.66***	-0.33***	-0.99***	-0.49***	-0.17***	-0.16***	-0.66***	-0.21***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[8]EST	-1.30***	-1.01***	-0.96***	-0.59***	0.16***	0.20***	-0.81***	-0.34***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[9]FRA	-0.80***	-0.66***	-0.36***	-0.16***	-0.11***	-0.12***	-0.58***	-0.32***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[10]HUN	-1.49***	-0.97***	-1.25***	-0.61***	0.0041	0.026	-0.88***	-0.33***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.890)	(0.429)	(0.000)	(0.000)
sh_country==[11]ITA	-0.54***	-0.39***	-1.48***	-1.27***	0.0061	-0.0024	-0.52***	-0.33***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.819)	(0.931)	(0.000)	(0.000)
sh_country==[12]NLD	-0.21***	-0.28***	0.32***	0.26***	-0.45***	-0.49***	0.12**	0.100**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)	(0.009)
sh_country==[13]POL	-0.72***	-0.23***	-0.95***	-0.37***	0.15***	0.18***	-0.96***	-0.44***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[14]PRT	-1.19***	-0.58***	-2.04***	-1.26***	0.10**	0.23***	-1.14***	-0.35***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)
sh_country==[15]SVN	-0.67***	-0.43***	0.066	0.38***	0.064*	0.074*	-0.30***	-0.069
	(0.000)	(0.000)	(0.114)	(0.000)	(0.047)	(0.034)	(0.000)	(0.120)
sh_country==[16]SWE	0.24***	0.079	0.012	-0.13**	0.31***	0.26***	0.12**	0.055
	(0.000)	(0.065)	(0.770)	(0.002)	(0.000)	(0.000)	(0.009)	(0.209)
Divorced/living separated		-0.11*		-0.047		-0.014		-0.12**
		(0.026)		(0.257)		(0.695)		(0.006)
Widowed		0.075		0.13**		0.10**		-0.15***
		(0.112)		(0.002)		(0.004)		(0.001)

Table A5.12 (continued): Regressing well-being and mental health on network types controlling for network size, relational dynamics and family status for all countries, all respondents with social support network.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
[1] Suburbs of big city		-0.0012		0.031		0.021		-0.090**
		(0.968)		(0.256)		(0.365)		(0.003)
[2] Large town		0.023		0.031		0.073***		-0.092***
		(0.413)		(0.223)		(0.001)		(0.001)
[3] Small town		0.089***		0.068**		0.062**		0.0053
		(0.001)		(0.004)		(0.001)		(0.838)
[4] Rural area/village		0.059*		0.072**		0.044*		-0.0075
		(0.021)		(0.002)		(0.020)		(0.764)
Employment, current job		0.17***		0.16***		0.011		0.099***
		(0.000)		(0.000)		(0.514)		(0.000)
Self-employment, current job		0.15***		0.16***		-0.033		0.077*
		(0.000)		(0.000)		(0.224)		(0.018)
[1] Primary school		0.10		0.37***		-0.051		0.24***
		(0.092)		(0.000)		(0.192)		(0.000)
[2] Lower secondary school		0.14*		0.47***		-0.049		0.33***
		(0.022)		(0.000)		(0.223)		(0.000)
[3] Upper secondary school		0.16**		0.56***		-0.067		0.43***
		(0.009)		(0.000)		(0.094)		(0.000)
[4] Post-secondary non-tertiary education		0.23**		0.66***		-0.041		0.50***
		(0.001)		(0.000)		(0.386)		(0.000)
[5] First stage tertiary education		0.23***		0.59***		-0.085*		0.43***
		(0.000)		(0.000)		(0.040)		(0.000)
[6] Second stage tertiary education		0.37***		0.70***		-0.091		0.41***
		(0.000)		(0.000)		(0.197)		(0.000)
[1] Fair health		1.01***		1.11***		0.11***		1.24***
		(0.000)		(0.000)		(0.000)		(0.000)
[2] Good health		1.49***		1.77***		0.12***		1.95***
		(0.000)		(0.000)		(0.000)		(0.000)
[3] Very good health		1.81***		2.13***		0.20***		2.26***
		(0.000)		(0.000)		(0.000)		(0.000)
[4] Excellent health		2.12***		2.44***		0.32***		2.38***
		(0.000)		(0.000)		(0.000)		(0.000)
Drugs for depression		-0.49***		-0.61***		-0.090***		-1.19***
		(0.000)		(0.000)		(0.000)		(0.000)
[1] Middle income		0.15***		0.19***		0.0018		0.082**
		(0.000)		(0.000)		(0.927)		(0.002)
[2] Upper middle income		0.23***		0.21***		0.010		0.068**
		(0.000)		(0.000)		(0.598)		(0.008)
[3] High income		0.23***		0.23***		-0.022		0.059*
		(0.000)		(0.000)		(0.215)		(0.016)
_cons	5.71***	2.78***	3.19***	-0.26	6.12***	6.05***	4.62***	2.12***
	(0.000)	(0.000)	(0.000)	(0.457)	(0.000)	(0.000)	(0.000)	(0.000)
N	50624	45591	48957	44214	50869	45770	50326	45316
R^2	0.14	0.25	0.21	0.38	0.16	0.17	0.11	0.31
adjusted R^2	0.14	0.25	0.21	0.38	0.16	0.17	0.11	0.31

Table A5.13: Regressing well-being and mental health on network types controlling for network size, relational dynamics and family status for all countries, male respondents with social support network.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
[2] Children	-0.11*	-0.049	-0.12*	-0.014	0.15***	0.15***	-0.10*	0.041
	(0.029)	(0.330)	(0.015)	(0.762)	(0.000)	(0.000)	(0.043)	(0.399)
[3] Other Relatives	-0.083	-0.039	0.023	0.084*	0.044	0.048	-0.11*	-0.040
	(0.090)	(0.405)	(0.621)	(0.049)	(0.222)	(0.207)	(0.026)	(0.370)
[4] Family	-0.055	0.013	-0.086*	0.0013	0.028	0.026	-0.074	0.030
	(0.161)	(0.734)	(0.020)	(0.969)	(0.297)	(0.356)	(0.052)	(0.400)
[5] Friends	0.038	0.038	0.15***	0.16***	0.050	0.057	-0.063	-0.033
	(0.411)	(0.395)	(0.000)	(0.000)	(0.158)	(0.123)	(0.164)	(0.428)
[6] Diverse	-0.042	-0.033	0.0076	0.048	-0.045	-0.061	-0.12*	-0.059
	(0.406)	(0.508)	(0.876)	(0.289)	(0.249)	(0.131)	(0.023)	(0.212)
Size of social network	0.077***	0.059***	0.084***	0.059***	0.064***	0.065***	0.041***	0.022*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.018)
Average contact 0-6	0.055**	0.053**	0.052**	0.051***	0.17***	0.16***	0.031	0.021
	(0.001)	(0.002)	(0.002)	(0.001)	(0.000)	(0.000)	(0.075)	(0.196)
Average closeness 0-3	0.30***	0.23***	0.30***	0.23***	0.70***	0.69***	0.17***	0.11***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average proximity 0-5	-0.045**	-0.0031	-0.076***	-0.028	-0.048***	-0.045***	-0.024	0.025
	(0.006)	(0.848)	(0.000)	(0.051)	(0.000)	(0.000)	(0.131)	(0.102)
Married/registered partnership	0.44***	0.37***	0.22***	0.14*	0.063**	0.043	0.22***	0.0011
	(0.000)	(0.000)	(0.000)	(0.016)	(0.009)	(0.385)	(0.000)	(0.984)
[1] Having 1 child	0.0044	-0.065	0.098*	-0.0076	-0.018	-0.036	-0.0053	-0.042
	(0.931)	(0.210)	(0.042)	(0.872)	(0.608)	(0.357)	(0.915)	(0.388)
[2] Having 2 children	0.16***	0.030	0.22***	0.068	-0.035	-0.049	0.10*	0.030
	(0.001)	(0.530)	(0.000)	(0.126)	(0.296)	(0.186)	(0.025)	(0.503)
[3] Having 3 or more children	0.092	-0.033	0.17***	0.0011	-0.046	-0.068	0.0093	-0.064
	(0.076)	(0.538)	(0.001)	(0.982)	(0.218)	(0.099)	(0.856)	(0.198)
Number of resident children	-0.010	-0.030	-0.090***	-0.11***	-0.033*	-0.038**	-0.014	-0.033
	(0.577)	(0.097)	(0.000)	(0.000)	(0.015)	(0.007)	(0.432)	(0.055)
Number of grandchildren	-0.0056	0.011*	-0.024***	-0.0025	0.0086*	0.010*	-0.015**	-0.0014
	(0.304)	(0.032)	(0.000)	(0.602)	(0.023)	(0.011)	(0.003)	(0.773)
Controls								
Age at interview	0.057***	0.084***	0.15***	0.16***	-0.016	-0.015	0.14***	0.12***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.165)	(0.236)	(0.000)	(0.000)
Age at interview, squared	-0.00039**	-0.00044***	-0.0012***	-0.0012***	0.00013	0.00014	-0.0012***	-0.00090***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.109)	(0.119)	(0.000)	(0.000)
sh_country==[2]BEL	-0.29***	-0.32***	-0.58***	-0.57***	-0.23***	-0.26***	-0.25***	-0.20***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[3]CHE	0.35***	0.055	0.46***	0.14**	0.26***	0.19***	0.15**	-0.11*
	(0.000)	(0.310)	(0.000)	(0.006)	(0.000)	(0.000)	(0.005)	(0.041)
sh_country==[4]CZE	-0.77***	-0.45***	-1.26***	-0.89***	-0.023	-0.0063	-0.053	0.27***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.561)	(0.891)	(0.316)	(0.000)
sh_country==[5]DEU	-0.34***	-0.29***	-0.076	-0.029	0.052	0.046	-0.0020	0.049
	(0.000)	(0.000)	(0.294)	(0.666)	(0.364)	(0.442)	(0.978)	(0.460)
sh_country==[6]DNK	0.36***	0.074	0.31***	-0.013	0.36***	0.29***	0.26***	0.047
	(0.000)	(0.189)	(0.000)	(0.793)	(0.000)	(0.000)	(0.000)	(0.378)
sh_country==[7]ESP	-0.52***	-0.27***	-0.78***	-0.37***	-0.14**	-0.12*	-0.18**	0.069
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.016)	(0.003)	(0.252)
sh_country==[8]EST	-1.42***	-1.15***	-1.09***	-0.72***	0.13**	0.13**	-0.75***	-0.35***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.005)	(0.000)	(0.000)
sh_country==[9]FRA	-0.75***	-0.65***	-0.28***	-0.16**	-0.089*	-0.10*	-0.41***	-0.26***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.027)	(0.017)	(0.000)	(0.000)
sh_country==[10]HUN	-1.45***	-0.97***	-1.21***	-0.61***	-0.023	-0.0067	-0.66***	-0.18**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.621)	(0.901)	(0.000)	(0.004)
sh_country==[11]ITA	-0.47***	-0.41***	-1.36***	-1.23***	0.032	0.025	-0.34***	-0.27***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.419)	(0.557)	(0.000)	(0.000)
sh_country==[12]NLD	-0.17***	-0.30***	0.32***	0.19***	-0.41***	-0.45***	0.27***	0.16**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)
sh_country==[13]POL	-0.64***	-0.19*	-0.90***	-0.34***	0.15**	0.18**	-0.75***	-0.29***
	(0.000)	(0.027)	(0.000)	(0.000)	(0.009)	(0.004)	(0.000)	(0.000)
sh_country==[14]PRT	-0.93***	-0.50***	-1.92***	-1.24***	0.14**	0.28***	-0.73***	-0.21**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.007)	(0.000)	(0.000)	(0.006)
sh_country==[15]SVN	-0.75***	-0.51***	0.081	0.39***	0.070	0.063	-0.22***	-0.0019
	(0.000)	(0.000)	(0.201)	(0.000)	(0.154)	(0.252)	(0.000)	(0.975)
sh_country==[16]SWE	0.22***	-0.012	-0.094	-0.29***	0.32***	0.26***	0.13*	0.014
	(0.000)	(0.848)	(0.132)	(0.000)	(0.000)	(0.000)	(0.035)	(0.814)
Divorced/living separated		-0.0057		0.033		-0.054		-0.100
		(0.938)		(0.605)		(0.342)		(0.128)
Widowed		0.094		0.12		0.020		-0.25**
		(0.242)		(0.078)		(0.750)		(0.001)

Table A5.13 (continued): Regressing well-being and mental health on network types controlling for network size, relational dynamics and family status for all countries, male respondents with social support network.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
[1] Suburbs of big city		-0.00046		0.034		0.033		-0.069
		(0.992)		(0.401)		(0.368)		(0.102)
[2] Large town		0.025		0.060		0.086*		-0.096*
		(0.553)		(0.115)		(0.011)		(0.015)
[3] Small town		0.074		0.092**		0.085**		0.028
		(0.060)		(0.008)		(0.006)		(0.430)
[4] Rural area/village		0.049		0.11**		0.096**		0.022
		(0.194)		(0.001)		(0.001)		(0.524)
Employment, current job		0.29***		0.21***		0.027		0.12***
		(0.000)		(0.000)		(0.297)		(0.000)
Self-employment, current job		0.21***		0.22***		0.0031		0.073
		(0.000)		(0.000)		(0.929)		(0.063)
[1] Primary school		0.24*		0.41***		-0.025		0.20*
		(0.012)		(0.000)		(0.721)		(0.026)
[2] Lower secondary school		0.25**		0.53***		0.022		0.29**
		(0.008)		(0.000)		(0.751)		(0.001)
[3] Upper secondary school		0.26**		0.61***		-0.0095		0.31***
		(0.006)		(0.000)		(0.891)		(0.001)
[4] Post-secondary non-tertiary education		0.37***		0.65***		0.057		0.30**
		(0.001)		(0.000)		(0.472)		(0.003)
[5] First stage tertiary education		0.30**		0.65***		-0.0024		0.29**
		(0.002)		(0.000)		(0.973)		(0.002)
[6] Second stage tertiary education		0.39**		0.72***		-0.089		0.23
		(0.004)		(0.000)		(0.393)		(0.079)
[1] Fair health		1.06***		1.13***		0.10**		1.22***
		(0.000)		(0.000)		(0.002)		(0.000)
[2] Good health		1.49***		1.76***		0.098**		1.86***
		(0.000)		(0.000)		(0.004)		(0.000)
[3] Very good health		1.81***		2.11***		0.19***		2.13***
		(0.000)		(0.000)		(0.000)		(0.000)
[4] Excellent health		2.08***		2.42***		0.30***		2.22***
		(0.000)		(0.000)		(0.000)		(0.000)
Drugs for depression		-0.43***		-0.58***		-0.081*		-1.21***
		(0.000)		(0.000)		(0.013)		(0.000)
[1] Middle income		0.23***		0.25***		0.026		0.14***
		(0.000)		(0.000)		(0.394)		(0.000)
[2] Upper middle income		0.30***		0.31***		0.021		0.17***
		(0.000)		(0.000)		(0.494)		(0.000)
[3] High income		0.26***		0.29***		0.038		0.12***
		(0.000)		(0.000)		(0.166)		(0.001)
_cons	4.79***	1.50*	2.37***	-1.00	6.84***	6.61***	3.71***	1.85**
	(0.000)	(0.014)	(0.000)	(0.062)	(0.000)	(0.000)	(0.000)	(0.001)
N	22018	19941	21359	19390	22121	20021	21882	19815
R^2	0.13	0.25	0.19	0.37	0.17	0.17	0.066	0.28
adjusted R^2	0.13	0.25	0.19	0.37	0.17	0.17	0.065	0.28

Table A5.14: Regressing well-being and mental health on network types controlling for network size, relational dynamics and family status for all countries, female respondents with social support network.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
[2] Children	-0.32***	-0.19***	-0.44***	-0.26***	0.21***	0.18***	-0.47***	-0.20***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
[3] Other Relatives	-0.20***	-0.11*	-0.17***	-0.063	0.20***	0.18***	-0.43***	-0.26***
	(0.000)	(0.019)	(0.000)	(0.138)	(0.000)	(0.000)	(0.000)	(0.000)
[4] Family	-0.17***	-0.082	-0.24***	-0.12**	0.18***	0.16***	-0.32***	-0.12**
	(0.000)	(0.052)	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)	(0.007)
[5] Friends	-0.16***	-0.11*	-0.069	0.028	0.25***	0.23***	-0.41***	-0.24***
	(0.001)	(0.020)	(0.111)	(0.502)	(0.000)	(0.000)	(0.000)	(0.000)
[6] Diverse	-0.26***	-0.11*	-0.28***	-0.093*	0.16***	0.14***	-0.57***	-0.31***
	(0.000)	(0.031)	(0.000)	(0.038)	(0.000)	(0.000)	(0.000)	(0.000)
Size of social network	0.11***	0.085***	0.12***	0.084***	0.075***	0.075***	0.060***	0.024**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)
Average contact 0-6	0.069***	0.075***	0.082***	0.078***	0.18***	0.18***	0.036*	0.039*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.033)	(0.012)
Average closeness 0-3	0.34***	0.29***	0.31***	0.26***	0.69***	0.67***	0.17***	0.12***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average proximity 0-5	-0.036**	-0.0026	-0.074***	-0.031**	-0.054***	-0.058***	-0.038**	0.017
	(0.007)	(0.847)	(0.000)	(0.009)	(0.000)	(0.000)	(0.006)	(0.204)
Married/registered partnership	0.43***	0.30***	0.19***	0.13**	-0.045**	0.016	0.15***	-0.057
	(0.000)	(0.000)	(0.000)	(0.008)	(0.008)	(0.704)	(0.000)	(0.314)
[1] Having 1 child	-0.091*	-0.088	0.024	0.020	0.041	0.022	-0.12*	-0.070
	(0.045)	(0.061)	(0.582)	(0.630)	(0.197)	(0.528)	(0.016)	(0.145)
[2] Having 2 children	0.051	0.045	0.094*	0.067	0.065*	0.042	0.023	0.056
	(0.239)	(0.322)	(0.025)	(0.110)	(0.034)	(0.217)	(0.629)	(0.235)
[3] Having 3 or more children	-0.022	-0.0067	0.019	0.018	0.063	0.040	-0.064	0.00077
	(0.659)	(0.897)	(0.694)	(0.699)	(0.064)	(0.291)	(0.232)	(0.988)
Number of resident children	-0.050*	-0.083***	-0.089***	-0.13***	-0.027*	-0.027*	-0.0047	-0.029
	(0.011)	(0.000)	(0.000)	(0.000)	(0.027)	(0.039)	(0.824)	(0.148)
Number of grandchildren	-0.0071	0.0032	-0.017***	-0.00072	0.0096**	0.0096**	-0.024***	-0.0072
	(0.157)	(0.513)	(0.000)	(0.863)	(0.002)	(0.003)	(0.000)	(0.138)
Controls								
Age at interview	0.0066	0.022	0.11***	0.12***	0.015	0.011	0.11***	0.11***
	(0.643)	(0.142)	(0.000)	(0.000)	(0.109)	(0.268)	(0.000)	(0.000)
Age at interview, squared	0.0000036	0.0000092	-0.00093***	-0.00090***	-0.000095	-0.000070	-0.00092***	-0.00071***
	(0.972)	(0.933)	(0.000)	(0.000)	(0.157)	(0.345)	(0.000)	(0.000)
sh_country==[2]BEL	-0.43***	-0.29***	-0.66***	-0.48***	-0.30***	-0.30***	-0.52***	-0.24***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[3]CHE	0.33***	0.14**	0.56***	0.33***	0.17***	0.19***	0.066	-0.12*
	(0.000)	(0.006)	(0.000)	(0.000)	(0.000)	(0.000)	(0.222)	(0.027)
sh_country==[4]CZE	-0.83***	-0.52***	-1.24***	-0.87***	-0.062	-0.074	-0.25***	0.093
	(0.000)	(0.000)	(0.000)	(0.000)	(0.066)	(0.053)	(0.000)	(0.065)
sh_country==[5]DEU	-0.31***	-0.25***	0.0034	0.099	0.079	0.086	-0.26***	-0.16*
	(0.000)	(0.000)	(0.958)	(0.107)	(0.119)	(0.109)	(0.000)	(0.017)
sh_country==[6]DNK	0.49***	0.27***	0.46***	0.21***	0.35***	0.34***	0.20**	-0.0061
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.919)
sh_country==[7]ESP	-0.78***	-0.38***	-1.16***	-0.60***	-0.19***	-0.19***	-1.03***	-0.44***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[8]EST	-1.22***	-0.90***	-0.87***	-0.49***	0.16***	0.24***	-0.85***	-0.35***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[9]FRA	-0.84***	-0.67***	-0.43***	-0.15***	-0.13***	-0.14***	-0.70***	-0.37***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
sh_country==[10]HUN	-1.52***	-0.96***	-1.27***	-0.61***	0.024	0.061	-1.05***	-0.41***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.525)	(0.148)	(0.000)	(0.000)
sh_country==[11]ITA	-0.60***	-0.39***	-1.57***	-1.30***	-0.013	-0.021	-0.66***	-0.37***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.711)	(0.570)	(0.000)	(0.000)
sh_country==[12]NLD	-0.25***	-0.26***	0.32***	0.32***	-0.49***	-0.51***	-0.0047	0.064
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.936)	(0.244)
sh_country==[13]POL	-0.79***	-0.26***	-1.00***	-0.40***	0.15**	0.19***	-1.12***	-0.54***
	(0.000)	(0.001)	(0.000)	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)
sh_country==[14]PRT	-1.39***	-0.64***	-2.14***	-1.27***	0.075	0.19***	-1.47***	-0.47***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.089)	(0.000)	(0.000)	(0.000)
sh_country==[15]SVN	-0.61***	-0.37***	0.050	0.38***	0.061	0.092*	-0.36***	-0.11
	(0.000)	(0.000)	(0.369)	(0.000)	(0.147)	(0.038)	(0.000)	(0.081)
sh_country==[16]SWE	0.24***	0.15**	0.084	-0.0020	0.28***	0.25***	0.11	0.086
	(0.000)	(0.009)	(0.134)	(0.970)	(0.000)	(0.000)	(0.089)	(0.159)
Divorced/living separated		-0.17**		-0.10		-0.00083		-0.13*
		(0.005)		(0.061)		(0.986)		(0.043)
Widowed		0.035		0.12*		0.11*		-0.12*
		(0.555)		(0.026)		(0.014)		(0.040)

Table A5.14 (continued): Regressing well-being and mental health on network types controlling for network size, relational dynamics and family status for all countries, female respondents with social support network.

	Life satisfaction		Quality of life (CASP-12)		Network satisfaction		Lack of depressive symptoms (EURO-D)	
	A	B	A	B	A	B	A	B
[1] Suburbs of big city		-0.0068		0.026		0.014		-0.11*
		(0.870)		(0.478)		(0.629)		(0.010)
[2] Large town		0.020		0.0073		0.062*		-0.092*
		(0.592)		(0.829)		(0.022)		(0.018)
[3] Small town		0.098**		0.044		0.047		-0.016
		(0.006)		(0.159)		(0.065)		(0.654)
[4] Rural area/village		0.065		0.039		0.0066		-0.033
		(0.065)		(0.199)		(0.784)		(0.344)
Employment, current job		0.085**		0.13***		-0.0057		0.071*
		(0.006)		(0.000)		(0.789)		(0.026)
Self-employment, current job		0.10*		0.088		-0.091*		0.093
		(0.044)		(0.072)		(0.034)		(0.102)
[1] Primary school		0.013		0.34***		-0.068		0.25**
		(0.869)		(0.000)		(0.155)		(0.001)
[2] Lower secondary school		0.053		0.41***		-0.10*		0.33***
		(0.493)		(0.000)		(0.042)		(0.000)
[3] Upper secondary school		0.084		0.50***		-0.11*		0.48***
		(0.281)		(0.000)		(0.030)		(0.000)
[4] Post-secondary non-tertiary education		0.12		0.65***		-0.11		0.62***
		(0.176)		(0.000)		(0.059)		(0.000)
[5] First stage tertiary education		0.19*		0.55***		-0.15**		0.53***
		(0.016)		(0.000)		(0.003)		(0.000)
[6] Second stage tertiary education		0.41**		0.68***		-0.039		0.58***
		(0.005)		(0.000)		(0.696)		(0.000)
[1] Fair health		0.98***		1.09***		0.11***		1.25***
		(0.000)		(0.000)		(0.000)		(0.000)
[2] Good health		1.50***		1.78***		0.13***		2.01***
		(0.000)		(0.000)		(0.000)		(0.000)
[3] Very good health		1.81***		2.15***		0.21***		2.37***
		(0.000)		(0.000)		(0.000)		(0.000)
[4] Excellent health		2.14***		2.45***		0.34***		2.50***
		(0.000)		(0.000)		(0.000)		(0.000)
Drugs for depression		-0.51***		-0.61***		-0.099***		-1.15***
		(0.000)		(0.000)		(0.000)		(0.000)
[1] Middle income		0.099**		0.14***		-0.011		0.051
		(0.005)		(0.000)		(0.656)		(0.172)
[2] Upper middle income		0.17***		0.13***		0.0029		0.0057
		(0.000)		(0.000)		(0.905)		(0.874)
[3] High income		0.21***		0.20***		-0.064**		0.035
		(0.000)		(0.000)		(0.007)		(0.302)
_cons	6.32***	3.68***	3.57***	0.12	5.75***	5.85***	4.65***	1.81***
	(0.000)	(0.000)	(0.000)	(0.795)	(0.000)	(0.000)	(0.000)	(0.001)
N	28606	25650	27598	24824	28748	25749	28444	25501
R^2	0.14	0.25	0.22	0.39	0.16	0.17	0.083	0.29
adjusted R^2	0.14	0.25	0.22	0.39	0.16	0.16	0.082	0.29

Chapter 6

Does Happiness Increase in Old Age? Longitudinal evidence from 20 European Countries

Abstract[**]

Several studies indicate that happiness follows a U-shape over the life cycle: Happiness decreases after the teenage years until reaching its nadir in middle age. A similar number of studies views the U-shape critically, stating that it is the result of the wrong controls or the wrong model. In this paper, we study the upward-pointing branch of the U-shape, tracing the happiness of European citizens 50 and older over multiple waves. Consistent with a U-shape around middle age, we find that happiness initially increases after the age of 50, but commonly stagnates afterwards and eventually reverts at high age. This pattern is generally observed irrespective of the utilized happiness measure, control variables, estimation methods, and the consideration of selection effects due to mortality. However, the strength of this pattern depends on the utilized happiness measure, control variables, and on mortality effects. The general pattern does not emerge for all countries, and is not always observed for women.

[**] Joint work with Stefan T. Trautmann. The content of this chapter has been published as: Becker, C. K., & Trautmann, S. T. (2022). Does Happiness Increase in Old Age? Longitudinal Evidence from 20 European Countries. Journal of Happiness Studies, 23(7), 3625-3654.

This paper uses data from SHARE Waves 1, 2, 4, 5, 6, and 7 (DOIs: 10.6103/SHARE.w1.710, 10.6103/SHARE.w2.710, 10.6103/SHARE.w4.710, 10.6103/SHARE.w5.710, 10.6103/SHARE.w6.710, 10.6103/SHARE.w7.711, see Börsch-Supan et al. (2013) for methodological details.

We thank Andrew Oswald and David Blanchflower for their helpful comments and insights.

The online supplement can be found here: `https://drive.google.com/drive/folders/16hW_5ZFJtRVrjDuCTQQCuYIMApHutb6W?usp=sharing`.

6.1 Introduction

Individual happiness can be gauged using various methods, for example self-reports of life satisfaction, measures of positive and negative affect or indirect measures, such as the number of antidepressants consumed. A substantial amount of work has been devoted to study how happiness measured in such ways develops over the course of the lifetime. This allows insight into how happiness evolves alongside important life events, such as changes in employment status, getting married, having children, but also ageing in general. Studies in economics often find that happiness decreases from the teenage years to middle age, only to increase afterwards (and then to fall again in very high age). This dip in middle age is referred to as the U-shape of happiness and has been reported for a variety of countries (Bell & Blanchflower, 2020; Blanchflower, 2021; Blanchflower & Graham, 2021; Blanchflower & Oswald, 2008; Piper, 2021; Gerdtham & Johannesson, 2001; Gwozdz & Sousa-Poza, 2010; Stone et al., 2010). This would indicate that people experience a low point of happiness around the age of 45-50. This dip is usually found to be comparable in magnitude to events such as getting divorced or losing employment (Blanchflower, 2021; Blanchflower & Graham, 2020). Taken together, this literature gives a persuasive reason to focus on this happiness dip as a researcher or policy maker. This is reflected in the attention this literature has received outside of academic research, reflected for example in articles in the Economist (2010) or the leading German weekly newspaper Die Zeit (2021), and many others.

At the same time, the U-shape around middle age has been contested by numerous other studies. Critique includes using the wrong controls (Glenn, 2009; Morgan & O'Connor, 2020), the wrong statistical model (Frijters & Beatton, 2012; Kratz & Brüderl, 2021; Ulloa et al., 2013), looking only at selected countries (Deaton 2008), neglecting sample attrition in panels caused by higher mortality among the unhappiest respondents (Hudomiet et al., 2021), and not accounting for cohort effects (Ulloa et al., 2013). This critique in turn has produced several replies, indicating that the U-shape exists, even when accounting for these critiques (Blanchflower & Graham, 2020; Blanchflower & Oswald, 2009; Clark, 2019). A further criticism is that a lot of evidence on the U-shape stems from cross-sectional data (Galambos et al., 2020; Ulloa et al., 2013), although some studies confirm the U-shape based on longitudinal data (Cheng et al., 2017; Clark & Oswald, 2006; Van Landeghem, 2012). Looking at cross-sections might produce a U-shape because events can affect disparate age groups differently. Crucially, there seems to be no clear consensus in the literature on which statistical tools should be used to estimate the relationship between age and happiness.

In this paper, we aim to add to this debate by providing an account based on a large European database. We use SHARE (Survey of Health, Age and Retirement in Europe) data, which includes people 50 and upwards. Accordingly, we study if happiness increases after middle age, the right branch of the U-shape. SHARE is a multi-wave panel; hence we add to the literature by providing further evidence for longitudinal data. We use different specifications and control sets based on previous literature to provide a detailed account of the age-happiness relation in old age for 20 European countries. Our results indicate support for a U-shape around middle age in the sense that happiness increases with age after midlife. Congruent with other studies we also find that happiness starts to deteriorate at high age (Blanchflower, 2021; Blanchflower & Graham, 2020; Gwozdz & Sousa-Poza, 2010). These results are generally robust to the specification used, as well as to using different subsets of the sample to account for country, gender, or selection effects due to mortality. Some countries do not or not clearly exhibit a positive relation between age and happiness. However, these results might in part be driven by lack of sufficient observations for the individual countries.

6.2 Methodology

6.2.1 Data

We use waves 1 to 7 of the SHARE Release 7.0.0 (Survey of Health, Age and Retirement in Europe) database (Börsch-Supan, 2018a,b,c,d,e,f; Börsch-Supan et al., 2013), except for wave 3. Wave 3 of SHARE (SHARELIFE) focused solely on past life events and does not include our target variables. SHARE is a database intended to be used to study the effects of aging over the life-course of European citizens aged 50 and older, managed by the Munich Center for the Economics of Aging, Max Planck Institute for Social Law and Social Policy. The cross-national panel database provides extensive data on health and socio-economic status. We merge data over the above-mentioned six waves in order to track respondents over the course of the different interviews. Respondents over the age of 80 were dropped due to small sample size. In total, the merged data set has 139,116 individual observations. Theses waves interviewed the respondents from 2004 to 2017, spanning 13 years and 20 countries. During this time some participants left the study (due to death or other reasons), while others joined (especially because later waves include additional countries).

6.2.2 Measuring Happiness

Measuring happiness, well-being or life satisfaction is crucial to our research question. How happy, well or satisfied people are with their life can depend on multiple domains, such as employment, relationships, physical and mental health, financial situation or the fulfillment of goals and desires (Easterlin, 1974; Frey & Stutzer, 2002). Accordingly, one can elicit broad measures of happiness (the simplest would be to ask respondents directly "How happy are you with your life?") or measures that zoom into specific domains. While there have been attempts to provide a unified, targetable index of happiness (such as Bhutan's Gross National Happiness or the Happy Planet Index), there is no consensus how to best measure happiness. In our study, we utilize three measures to map respondents' well-being: a simple single-item question regarding life satisfaction, the CASP-12 multi-item quality of life scale; and the EURO-D depressive symptoms scale. In the following, we discuss the three measures in more detail.

Our first measure, *life satisfaction*, measures a general, subjective feeling about the quality of life. It is extracted by a single-item question in which respondents indicate on a scale from 0 (low satisfaction) to 10 (high satisfaction) how satisfied they are with their life. This scale has acceptable reliability and validity (Pavot & Diener, 1993; Beckie & Hayduk, 1997) and relates meaningfully to various health and psychosocial measures (Kim et al., 2021).

Second, the CASP-12, a *quality of life* scale, which is designed to capture quality of life in old age (Hyde et al., 2003). Participants indicate for twelve statements whether they apply on a scale from 1 (often) to 4 (never). The twelve questions concern four dimensions of quality of life, control, autonomy, pleasure, and self-realization, resulting in an aggregate index ranging from 12 (low quality of life) to 48 (high quality of life). Hence, the CASP-12 relates more closely to affective measures or to the concept of *eudemonia*, where happiness follows from activity and control over one's life (see Aristoteles' Nicomachean Ethics, e.g. in Ameriks & Clarke 2000). We normalize it such that it ranges from 0 (low quality of life) to 10 (high quality of life).

Our third measure is the EURO-D depression score (Prince et al., 1999), which was designed to capture depressive symptoms among older people. It has been demonstrated to provide a valid comparison of depressive symptoms across European countries (Castro-Costa et al., 2008; Prince et al., 1999). The EURO-D depression score is generated from questions on 12 dimensions: Depression, pessimism, suicidality, guilt, sleep, interest, irritability, appetite, fatigue, concentration, enjoyment, and tearfulness. The answers to these questions result in an aggregate index ranging from 0 (not depressed) to 12 (very depressed). We normalize it such that it ranges from 0 (very depressed) to 10 (not depressed) and call it *lack of depressive symptoms*, such that higher values of this index are comparable to higher values in the other two measures.

Table A6.1 in the appendix provides an overview of the specific questions asked for these three measures. In the following sections, we address these three measures collectively as measures of *happiness*, unless specified otherwise.

6.2.3 Controls

Different events and choices in a person's life can influence the experienced level of happiness and life satisfaction (such as marrying, finding a better job, becoming a parent). If one wants to isolate the pure effect of aging on happiness, one might want to control for such factors. On the other hand, these events are an inherent part of aging. For example, many people become parents neither early nor very late in life. Controlling for such life events might thus lead to underestimating how happiness changes over the life course. If most of the important life events of a respondent are controlled for in their own variables, the effect of age is bound to become insignificant. As of yet, there appears to be no general agreement which set of controls should be included when analyzing happiness and life satisfaction in the literature.

Easterlin and Schaeffer (1999), Hellevik (2017) and Clark (2019) stressed the importance of controlling for cohort effects. Laaksonen (2018) showed that different controls sets can influence whether one obtains a U-shape (or any other specific form) in the first place. On the other hand, Frijters & Beatton (2012) favor fixed effects models that would exclude time-invariant controls, such as the birth cohort in order to account for unobserved heterogeneity. Finally, a number of studies (Blanchflower & Oswald, 2009; Blanchflower, 2021) have shown that the U-shape can be obtained even without using any controls at all. More importantly, if and which controls are used should depend on the underlying research question: Specifications *with* controls can capture the pure effect of aging, while abstracting from life events. Specifications *without* controls allow to estimate the overall trajectory of happiness over the life course (Blanchflower & Graham, 2020). A middle ground between those two approaches is to just include *exogenous* controls, that is only factors that remain constant over time, such as birth cohort or country of origin.

In order to accommodate the above outlined approaches, we conduct our analyses in the following ways: First, without any controls. Second, using a set of completely exogenous controls including gender, birth cohort, country of origin, and whether a respondent participated in all waves of the panel. Third, using a set of controls including the exogenous controls, as well as income, health and marriage/registered partnerships. Fourth, we use a fixed effect specification. One further concern can be the presence of selection effects: If less happy respondents are more likely to die early, they might disproportionally drop out of the panel, leading to a spurious positive correlation between age and happiness. Previous studies have indicated that different measures of happiness correlate positively with life expectation (Guven & Saloumidis, 2014; Lee & Singh, 2020; Kim et al., 2021). That is, older people could be happier, simply because their unhappier contemporaries are likely to die earlier and thus drop out of the pool of respondents. We control for this in three ways: First, we test whether we find evidence for such selection effects. Second, as we find such effects to be present (see section 6.3.2), we control for respondents that participated in all waves as mentioned above. This gives us a primary indication if selection effects might be present. Third, we conduct our main analysis for both, the full sample of all respondents and a subsample of respondents participating in all waves, thus excluding selection effects.

For the analyses including controls, we use the following variables from the SHARE data set as controls: Relationship status (1 if the respondent is married or in a registered partnership, 0 otherwise), gender (1 if female, 0 if male), age (of the respondent at the time of the interview), age squared, self-assessed physical health (measured on a 5-point scale from "poor" to "excellent"), and a dummy variable indicating the country of residence of the respondent to control for cultural differences. Further, we include the level of education according to the international classification of education ISCED-97 and brackets for the average monthly household income, which represent country-specific 25^{th}, 50^{th}, and 75^{th} percentiles of the reported household incomes from previous waves. This allows us to compare the effects of higher incomes across countries more easily. Additionally, we include a dummy variable for the birth cohort (which always covers a decade: 1930-1939, 1940-1949, etc.) and, as mentioned, a dummy variable for respondents that were present in all waves (subsequently called in all waves), to account for selection effects.

6.2.4 Models and Hypotheses

According to the previous sections, we estimate the following three models to test our research question. The observations of one participant in the different waves form a panel, standard errors are clustered on the level of the individual respondent.

$$M_{i,t} = \beta_0 + \beta D' + \gamma X'_{i,t} + u_{i,t} \tag{6.1}$$
$$M_{i,t} = \beta_0 + \beta_1 Age_{i,t} + \beta_2 Age^2_{i,t} + \gamma X'_{i,t} + u_{i,t} \tag{6.2}$$
$$M_{i,t} = \beta_0 + \beta_1 Age_{i,t} + \beta_2 Age^2_{i,t} + \gamma X'_{i,t} + \alpha_i + u_{i,t} \tag{6.3}$$

Equation (6.1) is a pooled OLS regression using dummies for different age categories, (6.2) is a pooled OLS regression including terms for age and ages squared. These two models are intended as a very basic test of a possible age-happiness relation, similar to Blanchflower & Oswald (2009). Equation (6.3) specifies a fixed effects GLS model, which allows to eliminate unobserved heterogeneity between respondents (Frijters & Beatton, 2012).[54] $M_{i,t}$ refers to our three happiness measures, life satisfaction, the CASP-12 index, and the EURO-D lack of depressive symptoms index, respectively (for individual $i = 1, ..., N$ and wave $t = 1, 2, 4, 5, 6, 7$). D is a vector of dummies quantifying age tuples starting from 52 (based on the literature of Blanchflower & Oswald 2009, 52-53, 54-55, and so on), respondents of younger age than that form the reference category (a total of 9,308 observations fall in this category). $Age_{i,t}$ and $Age^2_{i,t}$[55] refer to the age and the squared age of respondent i at time t. $X_{i,t}$ is a vector of time-varying (income brackets, education and subjective health) and time-invariant (gender, birth cohort and country of origin) personal controls (see section 6.2.3), α_i is the time-invariant personal effect of respondent i, and $u_{i,t}$ is an individual error term. As discussed in section 6.2.3, models are run without controls, with only the exogenous controls, or with all controls, the latter two control sets being represented by $X_{i,t}$.

All three model specifications test the same underlying research question: Does happiness increase after middle age (in line with the right side of the U-shape), after which it stagnates and eventually drops at high age? As our sample includes only respondents of age 50 and upwards, these two factors would imply a hill shaped path for the three happiness measures after middle age. Or put differently, a positive coefficient for age and a negative one for age squared (as happiness tends to fall for high age). In other words, we test:

Hypothesis 1: *The coefficients of the dummy variables β in model (6.1) are positive for lower ages, then close to zero and finally negative.*

[54] We use a simple fixed-effects specification here, comparable to Frijters & Beatton (2012) and in line with our simple pooled OLS specification. Fixed effects specifications can be sensitive to the baseline and still suffer from the identification problem, i.e., that age, time and cohort are perfectly collinear. There are attempts to rectify these problems, such as Van Landeghem (2012); Cheng et al. (2017); Dijk & Mierau (2018) and De Ree & Alessie (2011), wich go beyond the scope of this paper.

[55] In the fixed effects specification (6.3) time-invariant factors such as country of origin and or birth year are of course demeaned and thus eliminated from the estimation.

Hypothesis 2: *The age coefficients β_1 are positive and the age-squared coefficients β_2 are negative in models (6.2) and (6.3) (implying a concave shape, which would indicate that happiness increases after middle age and drops towards the end of life).*

Furthermore, we try to strengthen these hypotheses by running a series of robustness checks. First, as mentioned in section 6.2.3, one important concern studying happiness and old age is the presence of selection effects. In order to see if this concern is well-founded in our data set, we run the following fixed effects logit models:

$$Pr(Y_{i,t} = 1|x_{i,t}) = F(\beta_0 + \beta M_{i,t} + \gamma X'_{i,t} + \alpha_i + u_{i,t}) \tag{6.4}$$

Where $Pr(Y_{i,t} = 1|x_{i,t})$ is the probability that respondent i dies between wave t and wave t+1 ($Y = 1$), $M_{i,t}$ refers again to our three happiness measures, $X_{i,t}$ refers to the vector of control variables, α_i is the time-invariant personal effect of respondent i, and $u_{i,t}$ is an individual error term. If more happy people (according to our measures) are indeed less likely to die, we expect β to be negative. As discussed in section 6.2.3, we then take this into account for subsequent analyses. Additionally, our set of controls also contains the *in all waves* dummy variable. This allows us to capture any level effects caused by selection.

Second, we check whether our results differ if we perform some additional robustness checks. We run the regressions interacting the aforementioned *in all waves* dummy with the age and age squared variables. This provides further insight into the role of selection effects for the shape of the age-happiness relation. We also check if the age-happiness relation differs between male and female respondents, as well as between countries. Research has shown that the happiness of women and men differs (Laaksonen, 2018), and that the U-shape might be specific to some countries (Deaton, 2008). However, these control variables can only capture a level difference, not an overall different happiness-age pattern. Hence, we run our analyses again for men and women, as well as the different countries, separately.

6.3 Results

6.3.1 Summary Statisics

Table 6.1 provides an overview of key variables in our data set: the number of respondents per wave, percentages of female and married respondents, the average age, and our three variables on happiness and life satisfaction. The number of respondents increases over the waves, as further countries and more respondents were added. At the same time, other respondents dropped out of the survey due to attrition, noticeable in the drop in wave 7. Figure 6.1 provides an overview of the number of respondents per country. As visible, the number of respondents can vary considerably relative to the population size of the country. We account for this fact in our inferential analyses by using the sampling weights provided by SHARE.[56] Figure 6.2 shows the share of the various birth cohorts over the different waves, indicating that e.g. most respondents in the 1930-1939 birth cohort dropped out of the survey at one point. Figure 6.3 shows the number of living respondents relative to those that died before the wave was conducted, giving an overview of how the sample evolved over time. Respondents that do not drop out of the survey are interviewed again in subsequent waves, which overall leads to the average age of respondents increasing slightly over the waves.

Table 6.1 shows how the different measures for happiness and life satisfaction remained mostly stable on average over the waves. Before estimating the relationship between age and happiness, we can look at the raw answers to the different questions by age. Figure 6.4 shows the mean reported happiness over age pooled across all waves (see Figures S1-S6 in the online supplement for graphs for the individual waves). As the figure indicates, happiness seems indeed to increase with age starting from a low point in middle age in the raw data, before dropping strongly at high age.

Table 6.1: Summary statistics of key variables.

	Wave 1	Wave 2	Wave 4	Wave 5	Wave 6	Wave 7
N	23505	27478	45576	50488	50827	44961
Female	53%	54%	55%	54%	55%	55%
Married	76%	76%	72%	73%	72%	71%
Age	61.08	61.87	62.82	63.74	64.77	66.47
	(7.50)	(7.32)	(7.39)	(7.45)	(7.52)	(7.34)
Life satisfaction (0-10)	.	7.56	7.55	7.61	7.67	7.67
	.	(1.77)	(1.84)	(1.80)	(1.76)	(1.77)
CASP-12 (0-10)	7.06	7.02	7.05	7.32	7.07	7.11
	(1.68)	(1.71)	(1.76)	(1.69)	(1.72)	(1.73)
EURO-D (0-10)	8.12	8.15	7.92	8.11	8.05	8.10
	(1.85)	(1.87)	(1.88)	(1.82)	(1.85)	(1.85)

Notes: The values in rows four to seven report means, standard deviation in brackets.

[56] Sampling weights are inversely proportional to the probability of being sampled from the underlying population, based on demographic factors, such as nationality or gender. Sampling weights in SHARE are calculated using the procedure of (Deville & Särndal, 1992).

Figure 6.1: Number of respondents per country.

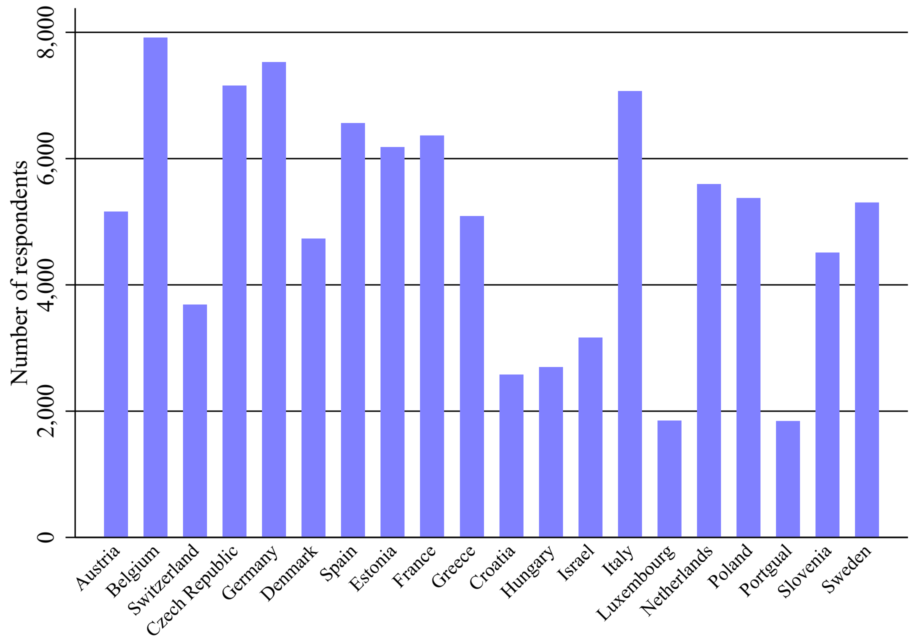


Figure 6.2: Distribution of birth cohorts in the different waves.

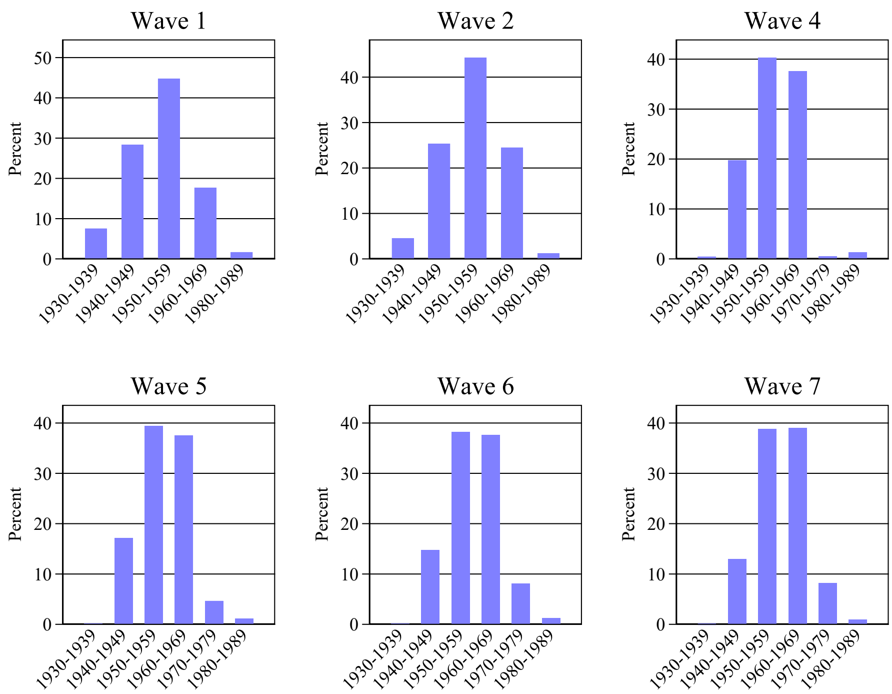


Figure 6.3: Number of living and deceased respondents in the different waves.

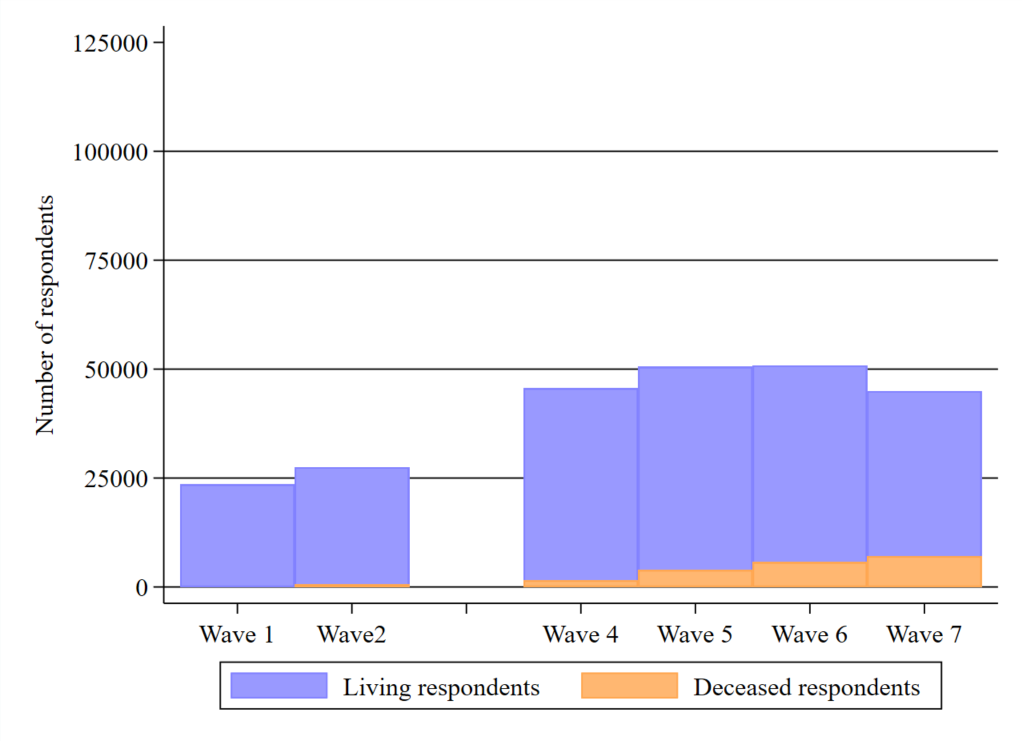
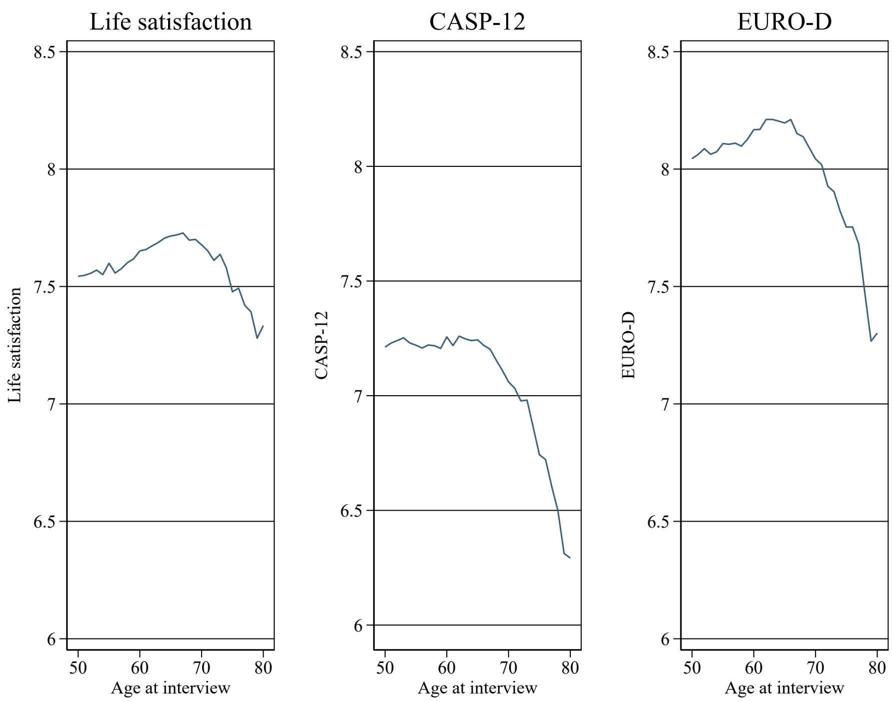


Figure 6.4: Raw values of the three happiness measures across all waves.

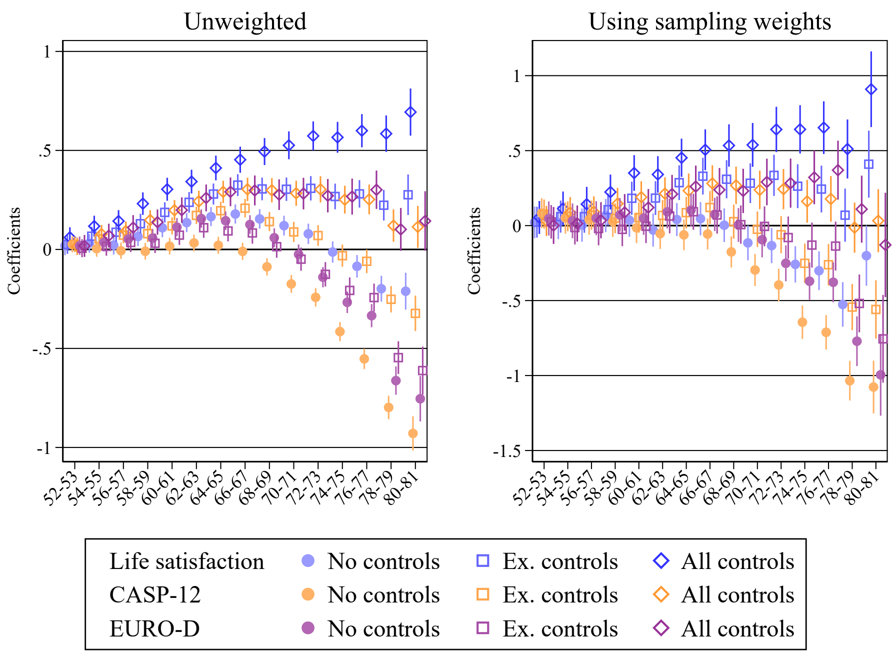


6.3.2 Main Analysis

Age and Happiness

Next, we estimate the relationship between the happiness measures and age. First, we are considering model (6.1), the pooled OLS model using age dummies. Figure 6.5 depicts the coefficients of the age dummies plotted against age for all respondents, with the left panel showing results without sampling weights and the right panel with weights. The implied happiness increases for all three measures starting with middle age, but tend to flatten or decrease in old age (the latter is a common finding in other studies, see e.g. Blanchflower & Oswald 2008; Deaton 2008). Including controls makes the increase after middle age even more pronounced, with the strong dip at old age becoming much less noticeable. A majority of the coefficients for the age dummies is highly significant (at $p < 0.001$, see tables S1-S6 in the online supplement for the full regressions) and follow the predicted path: Earlier age dummies are positive, while later ones are either negative or positive but smaller and ultimately not significant. An exception to this seems to be life satisfaction, once the full control set is included. As factors such as deteriorating health and changes in marriage status are accounted for, life satisfaction appears to increase over the course of life.

Figure 6.5: Coefficients and confidence intervals of the age dummies model (6.1).



Importantly, the effect sizes of the dummies (ranging from close to 0 to close to 1 at maximum) are similar to the results of other studies (Gwozdz & Sousa-Poza, 2010; Blanchflower & Graham, 2020; Blanchflower, 2021). To better illustrate the effect of ageing, we can also compare the effect sizes to that of important life events, such as getting divorced, losing a job, or losing a loved one (Blanchflower & Graham, 2020; Blanchflower, 2021). In our study for example, the effect of being married or in a registered partnership contributes between 0.171 and 0.411 to the happiness measures when using the full control set. Overall, these findings indicate a positive correlation between happiness and age with a tendency to flatten or decrease at high age, providing partial evidence in support of hypothesis 1.

Result 1: *The coefficients of the dummy variables β in model (6.1) are positive after middle age. Towards higher age they tend to become closer to zero or negative, depending on the model and control set used. Happiness increases with age but flattens of falls towards high age.*

These results are corroborated by the results of both the pooled OLS (6.2) and the fixed effects model (6.3) using age and age squared variables instead of dummies. Both models indicate an increase of all three measures over age that slows down, the older the respondents are. Table 6.2 displays the age and age squared coefficients of the pooled OLS model, again with and without sampling weights (the full regression tables are provided in tables S7-S12 in the online supplement[57]), Table 6.3 the ones of the fixed effects model (see tables S19-S21 in the online supplement for the full regressions).[58] As we test multiple hypotheses here on the same data set, a concern might be that the obtained significant results are suffering from multiple hypothesis testing. Tables 6.2 and 6.3 thus also display the t-statistics for the two models. As these statistics show, our results are highly significant. Furthermore, the results obtained from the fixed effects model are overall remarkably close to the ones from the pooled OLS. This would suggest, at least for our data, that using either model leads to valid results. In addition, Table 6.4 depicts the turning points implied by our models, i.e., the age where happiness starts to decrease. Generally, the more controls are included, the higher the turning points become. This is as expected, as controlling for changes e.g., in health or family status isolates negative shocks are more likely to occur in higher age. For CASP-12 and EURO-D this results in the turning point moving upwards, with this change never exceeding 10 years. For life satisfaction this effect is more pronounced. In fact, in the OLS models the turning point moves beyond the age range of our sample once all controls are included. Taken together, these results corroborate that the increase in happiness after middle age slows down and might ultimately turn into a decrease later in life. Overall, we thus find evidence for hypothesis 2.

Result 2: *The age coefficients β_1 are positive and significant and the age-squared coefficients β_2 are negative and significant in models (6.2) and (6.3). We find a concave shape for the age-happiness relation. Our results for CASP-12 and EURO-D indicate that happiness increases after middle age and drops towards the end of life. For life satisfaction, the drop becomes less pronounced as more controls are included.*

[57]Here we again run in the identification problem as age, cohort, and year are perfectly collinear and cannot be included simultaneously. Results using year dummies instead of cohort dummies are included in Table A6.2 in the appendix and in Tables S13-S18 in the online supplement. The estimates are overall qualitatively close to the cohort dummy specification.

[58]Longitudinal sampling weights for the fixed effects model require the respondents to be weighted over all waves. Hence, applying weights from wave 1 to wave 7 leads to all respondents that dropped out of the survey in between being dropped from the sample. The remaining sample is thus equal to the no attrition subsample and the weighted fixed effects results will be reported accordingly in section 6.3.2.

Table 6.2: Coefficients of the pooled OLS model (6.2) (using cohort dummies).

		Unweighted			Using sampling weights		
		No controls	Ex. controls	All controls	No controls	Ex. controls	All controls
Life satisfaction	Age	0.173***	0.118***	0.0915***	0.146***	0.122***	0.101***
		(17.28)	(10.51)	(6.92)	(6.57)	(4.86)	(3.74)
	Age^2	-0.00136***	-0.000829***	-0.000526***	-0.00123***	-0.000865***	-0.000582**
		(-17.39)	(-9.52)	(-5.10)	(-7.10)	(-4.39)	(-2.73)
CASP-12	Age	0.242***	0.191***	0.148***	0.250***	0.189***	0.154***
		(25.45)	(19.50)	(12.17)	(12.41)	(8.31)	(6.44)
	Age^2	-0.00204***	-0.00152***	-0.00108***	-0.00218***	-0.00159***	-0.00116**
		(-27.55)	(-19.85)	(-11.45)	(-13.99)	(-8.91)	(-6.22)
EURO-D	Age	0.274***	0.204***	0.138***	0.272***	0.171***	0.116***
		(26.59)	(17.74)	(8.88)	(11.58)	(6.24)	(3.73)
	Age^2	-0.00225***	-0.00169***	-0.00101***	-0.00227***	-0.00227***	-0.000835**
		(-27.77)	(-18.70)	(-8.34)	(-12.35)	(-6.55)	(-3.38)

Notes: t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 6.3: Coefficients of the fixed effects model (6.3).

		No controls	All controls
Life satisfaction	Age	0.154***	0.189***
		(12.79)	(10.70)
	Age^2	-0.00111***	-0.00129***
		(-11.80)	(-9.50)
CASP-12	Age	0.220***	0.216***
		(21.92)	(14.42)
	Age^2	-0.00172***	-0.00163***
		(-22.07)	(-14.17)
EURO-D	Age	0.255***	0.272***
		(19.40)	(11.42)
	Age^2	-0.00207***	-0.00196***
		(-20.13)	(-10.76)

Notes: t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 6.4: Age turning points of the OLS model (6.2) (using cohort dummies) and the FE model (6.3).

	OLS - Unweighted			OLS - Sampling weights			FE	
	No contr.	Ex. contr.	All contr.	No contr.	Ex. contr.	All contr.	No contr.	All contr.
Life satisfaction	63.60	71.17	86.98	59.35	70.52	86.77	69.37	73.26
CASP-12	59.31	62.83	68.52	57.34	59.43	66.38	63.95	66.26
EURO-D	60.89	60.36	68.32	59.91	60.64	69.46	61.59	69.39

Selection Effects

One major concern in the interpretation of the age effects shown in the previous sections is the presence of selection effects due to respondents dying depending on their happiness. Running fixed-effects logit regressions of the likelihood to die before a given wave on the different happiness measures (model (6.4) in section 6.2.4), we find indeed evidence of a selection effect. The regression coefficients for the three happiness measures are all negative and significant for the CASP-12 and EURO-D lack of depressive symptoms ($p < 0.05, p < 0.01$ for lack depressive symptoms, see Table A6.3 in the appendix). The likelihood of dying before a given wave decreases by 0.000126, 0.000189 and 0.000347 percentage points for each point on the scales of life satisfaction, CASP-12, and EURO-D lack of depressive symptoms, respectively. We additionally find that respondents with better physical health status are less likely to die. In the preceding section, the full control set also included the in all waves dummy for respondents that were present in all waves. The coefficients for this dummy are positive (Life satisfaction: 0.0412 [2.38], CASP: 0.104 [5.77], Lack of depressive symptoms: 0.0724 [3.68], t-statistics in square brackets, pooled OLS regression).

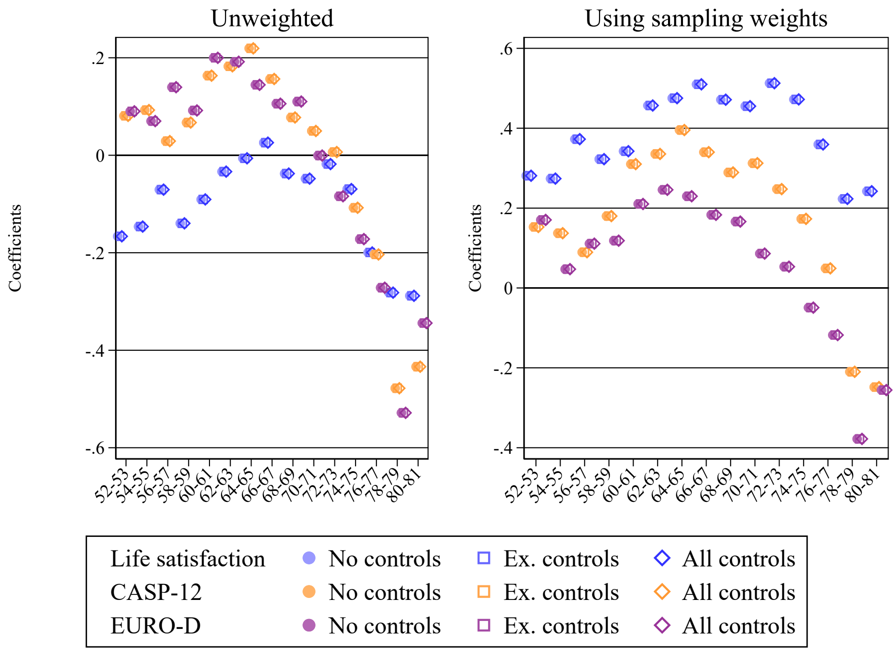
However, these coefficients can only account for a level effect between respondents that took part in all waves and those that dropped out of the sample at one point. To test if selection affects the shape of the happiness-age relation, we run our analyses for the subset of respondents that participated in all waves. Note that in the latter subset we also drop respondents that did not die between the waves, but either dropped out due to other reasons, or only joined the panel during the later waves. Past studies highlighted the fact that cross-sectional studies do not follow respondents over the life cycle and might thus have limited explanatory power (Ulloa et al., 2013; Galambos et al., 2020; Hudomiet et al., 2021): Accordingly, this subset represents the most stringent subset of respondents, specifically those for which we can track happiness over all waves.

Figure 6.6 shows the dummy coefficients of model (6.1) for the no attrition subset.[59] Looking at this subset, the obtained relationship between happiness and age emerges again, but loses part of its significance depending on the control set (in terms of the number of significant age dummies (see tables S22-S27 in the online supplement for the full regression). However, these results might in part be driven by the sharp decrease in observations once controls are used in the already strict no attrition subsample.

Tables 6.5 and 6.6 depict the age and age squared coefficients of the pooled OLS and fixed effects models (see tables S28-S36 in the online supplement for the full regressions). The estimated coefficients are all highly significant and fit our predictions. Comparing Tables 6.2 and 6.3 to Tables 6.5 and 6.6 shows that the coefficients are comparable in sign and size across the full sample and the no attrition subsample. We take this as further indication that selection effects are in place, but do not account for the observed correlation between happiness and age.

[59]Figure 6.6 shows the graphs without confidence intervals for better visibility. For the graph including the confidence intervals (which also illustrates the loss of significance), see Figure S7 in the online supplement.

Figure 6.6: Coefficients of the age dummies model (6.1), *no attrition* subsample.



Taking the findings of this section together, there is clear evidence that, while selection effects play a role, they seem to matter in the form of a level effect, rather than influencing the shape of the age-happiness relation. Notably, these results differ from the recent study of Hudomiet et al. (2021), which reports a decline in subjective well-being in U.S. data, as soon as attrition due to mortality is accounted for. Overall, our results are comparable, irrespective of whether we use sampling weights, account for attrition, using fixed effects, or using no or only exogenous controls. For the following subsample analysis, we hence use pooled OLS, sampling weights and the full set of controls for simplicity.

As a further robustness check, we run the weighted pooled OLS again, this time interacting the aforementioned in all waves dummy with the age and age squared variables (see Table A6.4 in the appendix and Table S37 in the online supplement). These interactions effects, as well as the in all waves dummy itself are in most cases insignificant. However, the coefficients for age and age squared still exhibit the same pattern in our main analysis. A notable exception is the CASP-12: Including the interaction effects here renders the in all waves dummy itself significant, but negative. The interaction effects with age and age squared are significant, and are also positive and negative, respectively. In other words, even in this exception, respondents that took part in all waves exhibit the same age-happiness pattern as in the main analysis. If anything, the pattern emerges even stronger here.

Table 6.5: Coefficients of the pooled OLS model (6.2) (using cohort dummies), no attrition subsample.

		Unweighted			Using sampling weights		
		No controls	Ex. controls	All controls	No controls	Ex. controls	All controls
Life satisfaction	Age	0.177^{***}	0.154^{***}	0.153^{**}	0.200^{***}	0.192^{***}	0.191^{**}
		(5.01)	(4.90)	(2.74)	(4.62)	(5.31)	(2.82)
	Age^2	-0.00134^{***}	-0.00114^{***}	-0.00104^{*}	-0.00147^{***}	-0.00142^{***}	-0.00129^{**}
		(-5.00)	(-4.74)	(-2.54)	(-4.49)	(-5.11)	(-2.59)
CASP-12	Age	0.253^{***}	0.233^{***}	0.327^{***}	0.305^{***}	0.257^{***}	0.343^{***}
		(8.10)	(9.03)	(5.83)	(8.13)	(8.40)	(4.88)
	Age^2	-0.00203^{***}	-0.00173^{***}	-0.00232^{***}	-0.00235^{***}	-0.00190^{***}	-0.00240^{**}
		(-8.36)	(-8.68)	(-5.61)	(-8.12)	(-7.97)	(-4.66)
EURO-D	Age	0.236^{***}	0.201^{***}	0.149^{*}	0.231^{***}	0.201^{***}	0.166^{*}
		(7.58)	(7.34)	(2.35)	(5.96)	(6.12)	(2.18)
	Age^2	-0.00193^{***}	-0.00165^{***}	-0.00113^{*}	-0.00184^{***}	-0.00163^{***}	-0.00123^{*}
		(-8.00)	(-7.76)	(-2.42)	(-6.18)	(-6.39)	(-2.20)

Notes: t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 6.6: Coefficients of the fixed effects model (6.3), no attrition subsample.

		No controls	All controls
Life satisfaction	Age	0.171^{***}	0.204^{***}
		(5.72)	(3.50)
	Age^2	-0.00128^{***}	-0.00154^{***}
		(-5.61)	(-3.60)
CASP-12	Age	0.204^{***}	0.357^{***}
		(7.83)	(6.24)
	Age^2	-0.00143^{***}	-0.00250^{***}
		(-7.13)	(-5.95)
EURO-D	Age	0.243^{***}	0.240^{***}
		(7.85)	(3.32)
	Age^2	-0.00197^{***}	-0.00184^{***}
		(-8.28)	(-3.46)

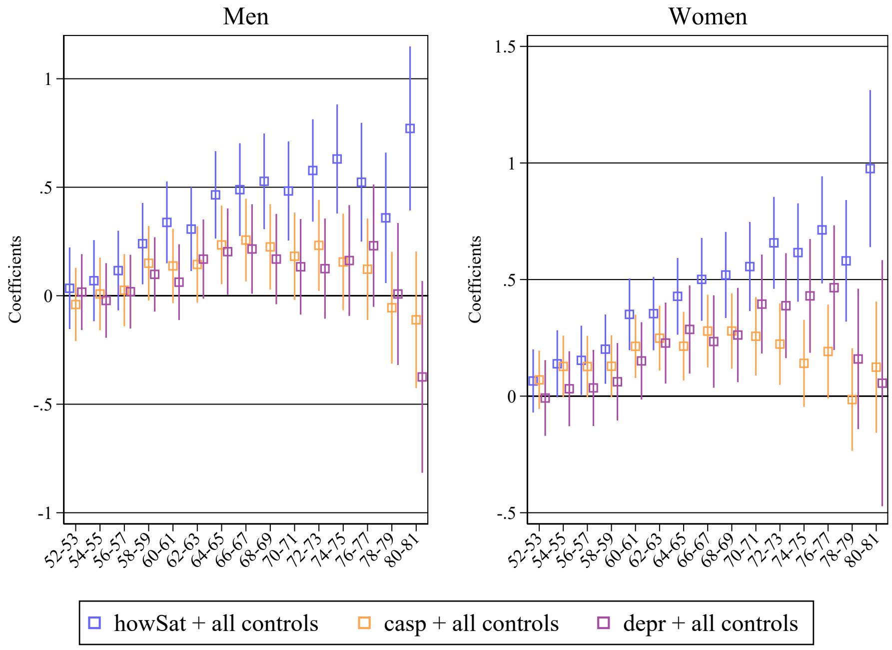
Notes: t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

6.3.3 Subsample Analyses

Gender Differences

Looking at men and women separately, the results of the dummy regressions in Figure 6.7 already indicate that the age-happiness relation follows a comparable path for both genders (see tables S38-S39 in the online supplement for the full regressions). Table 6.7 shows the coefficients and t-statistics for the pooled OLS model (6.1) (see tables S40-S41 in the online supplement for the full regressions). In general, the results look similar for both men and women. However, for women, the pattern is less pronounced, falling just short of reaching significance for some of the coefficients in the pooled OLS model, except for the CASP-12. We run the same regression with interaction terms (see Table A6.5 in the appendix and Table S42 in the online supplement). None of the interaction terms are significant, corroborating that the fundamental pattern is similar for men and women.

Figure 6.7: Coefficients and confidence intervals of the age dummies model (6.1) for men and women, all respondents.



Country Differences

Next, we turn to the differences between the countries of the SHARE data set. For the dummy regression plots for the 20 individual countries, see Figure A6.1 in the appendix (full dummy regressions in tables S43-S62 in the online supplement). For an overview of all age and age squared coefficients of the pooled OLS how to best measure happiness best model, see Table A6.6 in the appendix (full dummy regressions in tables S63-S82 in the online supplement). Evidence from the pooled OLS model here is mixed, with some countries not observing a significant correlation

Table 6.7: Coefficients of the pooled OLS model (6.2), men and women.

		Men	Women
Life satisfaction	Age	0.153^{***}	0.0681
		(3.79)	(1.88)
	Age^2	-0.001000^{**}	-0.000319
		(-3.16)	(-1.12)
CASP-12	Age	0.183^{***}	0.135^{***}
		(5.11)	(4.24)
	Age^2	-0.00138^{***}	-0.00102^{***}
		(-4.96)	(-4.10)
EURO-D	Age	0.132^{*}	0.0959^{*}
		(2.99)	(2.24)
	Age^2	-0.00100^{**}	-0.000648
		(-2.84)	(-1.91)

Notes: t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

between age and happiness at all (or only for some of the happiness measures used). Still, for all countries and measures for which a significant correlation is observed, the positive trend for happiness with age and the negative with age squared is obtained. Notably, however, the pooled OLS with age coefficients and pooled OLS dummy regressions do not always agree in terms of the significance level. Belgium (panel 2 in Figure A6) for example exhibits a positive relation between age and happiness for life satisfaction and the CASP-12, while the corresponding coefficients in the pooled OLS regression fail to reach significance.

Of course, conducting the analysis for each country separately with the full control set additionally atomises the data. This is only exacerbated by different countries having differing sample sizes in the data set to begin with. As measures such as the question on life satisfaction appear in many questionnaires, our results could be complemented by studying larger national data sets. Alternatively, future waves of SHARE might include further data to answer the question if the observed insignificances are caused here by a lack of data points or by some countries not exhibiting a positive relation between age and happiness.

6.4 Discussion

Studies measuring happiness and well-being over the life cycle have found mixed results, and in particular the U-shape of happiness is a controversial finding. Consistent with a U-shape around middle age, we find that happiness increases after the age of 50, irrespective of the specification used. Furthermore, our results indicate that happiness tends to stagnate or even decrease at very high age. When conducting our analysis on country- or gender-specific subsamples, a more varied picture emerges. Where we find significant results in these subsamples, however, it is always consistent with a U-shape. These findings are also robust when accounting for differences due to mortality selection effects. While selection effects are indeed at work, with happier respondents being more likely to be alive at the time the next wave is elicited, CASP-12 is the only measure where the pattern is affected: selection makes the observed pattern more pronounced in this case. The result could potentially stem from the CASP-12 measuring control and agency, which decrease towards the end of one's life (Ribeiro et al., 2020; Rodríguez-Blázquez et al., 2020; Oliver et al., 2021). This might also help to explain why we find lower turning points for CASP-12 and EURO-D in Table 6.4 in contrast to life satisfaction, when including additional controls. One reason why life satisfaction might continue to increase in high age is that older people might give up on aspiration and enjoy life more (Blanchflower & Oswald, 2004; Frey & Stutzer, 2010). CASP-12 and EURO-D, on the other hand, measure elements related to control and mental health, which might be more negatively affected by age. Different happiness measures might capture different aspects of life, highlighting the importance of looking at multiple measures at the same time.

Importantly, the observed age-happiness relation is consistently obtained using different approaches that have been used in both research that found and did not find the happiness dip in middle age. Additionally, the happiness-age relationship does not only hold for measures of subjective well-being (life satisfaction), but also for affective/eudemonic (CASP-12) and mental health measures (EURO-D). We are thus confident that our findings are meaningful for a substantial number of European countries.

Naturally, we can make no predictions about the trajectory of the happiness-age relation under the age of 50, as the SHARE data set only provides data for older Europeans. However, as other studies have indicated, there is support for the overall U-shape in various European countries (Blanchflower, 2021). We find that happiness indeed increases after middle age, compared to other studies finding a decrease after middle age (Mroczek & Spiro, 2005; Easterlin, 2006) or an overall decrease (Frijters & Beatton, 2012; Kassenboehmer & Haisken-DeNew, 2012). These differences could reflect regional differences, as Easterlin (2006) and Mroczek & Spiro (2005) use US data. Alternatively, methodological differences might drive these divergences. Kassenboehmer & Haisken-DeNew (2012) utilize respondents leaving the survey panel temporarily, to differentiate between age and years in the survey. Both should still be correlated, however. Frijters & Beatton (2012) main result is based on fixed effects regressions, which might ultimately not be reliable enough to deal with the age-period-cohort problem (Heckman & Robb Jr, 1985; Yang & Land, 2008). Mrozcek and Spiro's (2005) use of a demeaned variable in their specification might similarly be problematic (McIntosh & Schlenker, 2006).

Our results are in line with previous studies indicating an increase of happiness after 50 (Morgan & O'Connor, 2017) or an upward profile for affective measures (Mroczek & Kolarz, 1998). However, similar to other studies, our results also provide evidence that happiness, depending on the measure used, stagnates or even decreases later in life (Gwozdz & Sousa-Poza, 2010; Blanchflower & Graham, 2020; Blanchflower, 2021). Our results support the view that people go through a period of relatively low happiness (relative to happiness at older age) around the midpoint of their life. For policy makers, it is important to further explore why this dip occurs and how it can be alleviated.

Going forward, it is important to highlight that proving or disproving the U-shape of happiness, or as in our case components of it, should not be a goal in itself. While knowing the average path happiness takes over the course of a human life is important, even more so is understanding which life events affect the emerging trajectory (Bjørnskov et al., 2008; Lachman, 2015; Morgan & O'Connor, 2020; Galambos et al., 2020, 2021). Past research has shown the happiness effects of marriage (Grover & Helliwell, 2019), parenthood (Nelson et al., 2013), social networks in general (Becker et al., 2019), income (Easterlin, 1974), social support (Siedlecki et al., 2014), permanent employment (Piper, 2021), the quality of formal institutions (Bjørnskov et al., 2010), giving up on aspirations (Schwandt, 2016), and health (Gwozdz & Sousa-Poza, 2010; Oliver et al., 2021; Bussière et al., 2021). Mapping the evolution of these events over the life course may help to better understand the emergence of the U-shape of happiness.

Chapter 6 References

Ameriks, K., & Clarke, D. M. (2000). *Aristotle: Nicomachean Ethics*. Cambridge University Press.

Becker, C., Kirchmaier, I., & Trautmann, S. T. (2019). Marriage, parenthood and social network: Subjective well-being and mental health in old age. *PloS one*, *14*(7), e0218704. (Publisher: Public Library of Science San Francisco, CA USA)

Beckie, T. M., & Hayduk, L. A. (1997). Measuring Quality of Life. *Social Indicators Research*, *42*, 21–39.

Bell, D. N., & Blanchflower, D. G. (2020). The U-shape of Happiness in Scotland. *National Bureau of Economic Research*(No. w28144).

Bjørnskov, C., Dreher, A., & Fischer, J. A. (2008). Cross-country determinants of life satisfaction: Exploring different determinants across groups in society. *Social Choice and Welfare*, *30*(1), 119–173. (Publisher: Springer)

Bjørnskov, C., Dreher, A., & Fischer, J. A. (2010). Formal institutions and subjective well-being: Revisiting the cross-country evidence. *European Journal of Political Economy*, *26*(4), 419–430. (Publisher: Elsevier)

Blanchflower, D. G. (2021). Is happiness U-shaped everywhere? Age and subjective well-being in 145 countries. *Journal of Population Economics*, *34*(2), 575–624. (Publisher: Springer)

Blanchflower, D. G., & Graham, C. L. (2020). The mid-life dip in well-being: Economists (who find it) versus psychologists (who don't)!

Blanchflower, D. G., & Graham, C. L. (2021). The U-shape of happiness: A response. *Perspectives on Psychological Science*, 1745691620984393. (Publisher: SAGE Publications Sage CA: Los Angeles, CA)

Blanchflower, D. G., & Oswald, A. J. (2004). Well-being over time in britain and the usa. *Journal of public economics*, *88*(7-8), 1359–1386.

Blanchflower, D. G., & Oswald, A. J. (2008). Is well-being U-shaped over the life cycle? *Social Science & Medicine*, *66*(8), 1733–1749.

Blanchflower, D. G., & Oswald, A. J. (2009). The U-shape without controls: A response to Glenn. *Social Science & Medicine*, *69*(4), 486–488. (Publisher: Elsevier)

Bussière, C., Sirven, N., & Tessier, P. (2021). Does ageing alter the contribution of health to subjective well-being? *Social Science & Medicine*, *268*(113456). (Publisher: Elsevier)

Börsch-Supan, A. (2018a). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 1. Release version: 7.0.0. *SHARE-ERIC*.

Börsch-Supan, A. (2018b). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 2. Release version: 70.0. *SHARE-ERIC*.

Börsch-Supan, A. (2018c). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4. Release version: 7.0.0. *SHARE-ERIC*.

Börsch-Supan, A. (2018d). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5. Release version: 7.0.0. *SHARE-ERIC*.

Börsch-Supan, A. (2018e). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 6. Release version: 7.0.0. *SHARE-ERIC*.

Börsch-Supan, A. (2018f). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 7. Release version: 7.0.0. *SHARE-ERIC*.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., ... Zuber, S. (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, *42*(4), 992–1001.

Castro-Costa, E., Dewey, M., Stewart, R., Banerjee, S., Huppert, F., Mendonca-Lima, C., ... Prince, M. (2008, March). Ascertaining late-life depressive symptoms in Europe: an evaluation of the survey version of the EURO-D scale in 10 nations. The SHARE project. *International Journal of Methods in Psychiatric Research*, *17*(1), 12–29.

Cheng, T. C., Powdthavee, N., & Oswald, A. J. (2017). Longitudinal evidence for a midlife nadir in human well-being: Results from four data sets. *The Economic Journal*, *127*(599), 126–142. (Publisher: Oxford University Press Oxford, UK)

Clark, A. E. (2019). Born to be mild? Cohort effects don't (fully) explain why well-being is U-shaped in age. In *The Economics of Happiness* (pp. 387–408). Springer.

Clark, A. E., & Oswald, A. J. (2006). The curved relationship between subjective well-being and age. *halshs-00590404*.

Deaton, A. (2008). Income, health, and well-being around the world: Evidence from the Gallup World Poll. *Journal of Economic Perspectives*, *22*(2), 53–72.

De Ree, J., & Alessie, R. (2011). Life satisfaction and age: Dealing with underidentification in age-period-cohort models. *Social Science & Medicine*, *73*(1), 177–182. (Publisher: Elsevier)

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, *87*(418), 376–382. (Publisher: Taylor & Francis)

Dijk, H. H., & Mierau, J. O. (2018). Mental Health over the Life Course: Evidence for a U-Shape? *European Journal of Public Health*, *28*(suppl_4), cky213–885. (Publisher: Oxford University Press)

Easterlin, R. A. (1974). Does economic growth improve the human lot? Some empirical evidence. In *Nations and households in economic growth* (pp. 89–125). Elsevier.

Easterlin, R. A. (2006). Life cycle happiness and its sources: Intersections of psychology, economics, and demography. *Journal of Economic Psychology*, *27*(4), 463–482. (Publisher: Elsevier)

Economist. (2010). The U-bend of life. Retrieved from `https://www.economist.com/christmas-specials/2010/12/16/the-u-bend-of-life`

Frey, B. S., & Stutzer, A. (2002). What can economists learn from happiness research? *Journal of Economic Literature*, *40*(2), 402–435.

Frey, B. S., & Stutzer, A. (2010). Happiness and economics. In *Happiness and Economics.* Princeton University Press.

Frijters, P., & Beatton, T. (2012). The mystery of the U-shaped relationship between happiness and age. *Journal of Economic Behavior & Organization*, *82*(2-3).

Galambos, N. L., Krahn, H. J., Johnson, M. D., & Lachman, M. E. (2020). The U shape of happiness across the life course: Expanding the discussion. *Perspectives on Psychological Science*, *15*(4), 898–912. (Publisher: SAGE Publications Sage CA: Los Angeles, CA)

Galambos, N. L., Krahn, H. J., Johnson, M. D., & Lachman, M. E. (2021). Another attempt to move beyond the cross-sectional U shape of happiness: A reply. *Perspectives on psychological science*, *16*(6), 1447–1455. (Publisher: SAGE Publications Sage CA: Los Angeles, CA)

Gerdtham, U.-G., & Johannesson, M. (2001). The relationship between happiness, health, and socio-economic factors: results based on Swedish microdata. *The Journal of Socio-Economics*, *30*(6), 553–557.

Glenn, N. (2009). Is the apparent U-shape of well-being over the life course a result of inappropriate use of control variables? A commentary on Blanchflower and Oswald (66: 8, 2008, 1733–1749). *Social Science & Medicine*, *69*(4), 481–485.

Grover, S., & Helliwell, J. F. (2019). How's life at home? New evidence on marriage and the set point for happiness. *Journal of Happiness Studies*, *20*(2), 373–390. (Publisher: Springer)

Guven, C., & Saloumidis, R. (2014). Life satisfaction and longevity: Longitudinal evidence from the German Socio-Economic Panel. *German Economic Review*, *15*(4), 453–472. (Publisher: De Gruyter)

Gwozdz, W., & Sousa-Poza, A. (2010). Ageing, health and life satisfaction of the oldest old: An analysis for Germany. *Social Indicators Research*, *97*(3), 397–417. (Publisher: Springer)

Heckman, J. J., & Robb Jr, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of econometrics*, *30*(1-2), 239–267.

Hudomiet, P., Hurd, M. D., & Rohwedder, S. (2021). The age profile of life satisfaction after age 65 in the US. *Journal of Economic Behavior & Organization*, *189*, 431–442. (Publisher: Elsevier)

Hyde, M., Wiggins, R. D., Higgs, P., & Blane, D. B. (2003, May). A measure of quality of life in early old age: The theory, development and properties of a needs satisfaction model (CASP-19). *Aging & Mental Health*, *7*(3), 186–194.

Kassenboehmer, S. C., & Haisken-DeNew, J. P. (2012). Heresy or enlightenment? The well-being age U-shape effect is flat. *Economics Letters*, *117*(1), 235–238. (Publisher: Elsevier)

Kim, E. S., Delaney, S. W., Tay, L., Chen, Y., Diener, E. D., & Vanderweele, T. J. (2021). Life satisfaction and subsequent physical, behavioral, and psychosocial health in older adults. *The Milbank Quarterly*, *99*(1), 209–239. (Publisher: Wiley Online Library)

Kratz, F., & Brüderl, J. (2021). The age trajectory of happiness. (Publisher: PsyArXiv)

Laaksonen, S. (2018). A research note: Happiness by age is more complex than u-shaped. *Journal of Happiness Studies*, *19*, 471–482.

Lachman, M. E. (2015). Mind the gap in the middle: A call to study midlife. *Research in Human Development*, *12*(3-4), 327–334. (Publisher: Taylor & Francis)

Lee, H., & Singh, G. K. (2020). Inequalities in Life Expectancy and All-Cause Mortality in the United States by Levels of Happiness and Life Satisfaction: A Longitudinal Study. *International Journal of Maternal and Child Health and AIDS*, *9*(3), 305. (Publisher: Global Health and Education Projects, Inc.)

McIntosh, C. T., & Schlenker, W. (2006). Identifying non-linearities in fixed effects models. *UC-San Diego Working Paper*.

Morgan, R., & O'Connor, K. J. (2017). Experienced life cycle satisfaction in Europe. *Review of Behavioral Economics*, *4*(4), 371–396. (Publisher: Now Publishers, Inc.)

Morgan, R., & O'Connor, K. J. (2020). Does the U-shape Pattern in Life Cycle Satisfaction Obscure Reality? A Response to Blanchflower. *Review of Behavioral Economics*, *7*(2), 201–206. (Publisher: now publishers)

Mroczek, D. K., & Kolarz, C. M. (1998). The effect of age on positive and negative affect: a developmental perspective on happiness. *Journal of Personality and Social Psychology*, *75*(5), 1333. (Publisher: American Psychological Association)

Mroczek, D. K., & Spiro, A. (2005). Change in life satisfaction during adulthood: findings from the veterans affairs normative aging study. *Journal of Personality and Social Psychology*, *88*(1), 189. (Publisher: American Psychological Association)

Nelson, S. K., Kushlev, K., English, T., Dunn, E. W., & Lyubomirsky, S. (2013, January). In Defense of Parenthood: Children Are Associated With More Joy Than Misery. *Psychological Science*, *24*(1), 3–10.

Novotny, R. (2021). Älter werden: Das Beste kommt noch. *ZEIT*. Retrieved from `https://www.zeit.de/2021/05/aelter-werden-glueck-zufriedenheit-leben-philosophie`

Oliver, A., Sentandreu-Mañó, T., Tomás, J. M., Fernández, I., & Sancho, P. (2021). Quality of life in European older adults of SHARE wave 7: Comparing the old and the oldest-old. *Journal of Clinical Medicine*, *10*(13), 2850. (Publisher: MDPI)

Pavot, W., & Diener, E. (1993). The Affective and Cognitive Context of Self-Reported Measures of Subjective Well-Being. *Social Indicators Research*, *28*(1), 1–20.

Piper, A. (2021). Temps dip deeper: Temporary employment and the midlife nadir in human well-being. *The Journal of the Economics of Ageing*, *19*, 100323. (Publisher: Elsevier)

Prince, M. J., Reischies, F., Beekman, A. T. F., Fuhrer, R., Jonker, C., Kivela, S. L., ... Copeland, J. R. M. (1999). Development of the EURO–D scale – a European Union initiative to compare symptoms of depression in 14 European centres. *British Journal of Psychiatry*, *174*(04), 330–338.

Ribeiro, O., Teixeira, L., Araújo, L., Rodríguez-Blázquez, C., Calderón-Larrañaga, A., & Forjaz, M. J. (2020). Anxiety, depression and quality of life in older adults: Trajectories of influence across age. *International journal of environmental research and public health*, *17*(23), 9039. (Publisher: MDPI)

Rodríguez-Blázquez, C., Ribeiro, O., Ayala, A., Teixeira, L., Araújo, L., & Forjaz, M. J. (2020). Psychometric properties of the CASP-12 scale in Portugal: An analysis using SHARE data. *International Journal of Environmental Research and Public Health*, *17*(18), 6610. (Publisher: MDPI)

Schwandt, H. (2016). Unmet aspirations as an explanation for the age U-shape in wellbeing. *Journal of Economic Behavior & Organization*, *122*, 75–87. (Publisher: Elsevier)

Siedlecki, K. L., Salthouse, T. A., Oishi, S., & Jeswani, S. (2014). The relationship between social support and subjective well-being across age. *Social indicators research*, *117*(2), 561–576. (Publisher: Springer)

Stone, A. A., Schwartz, J. E., Broderick, J. E., & Deaton, A. (2010). A snapshot of the age distribution of psychological well-being in the United States. *Proceedings of the National Academy of Sciences*, *107*(22), 9985–9990. (Publisher: National Acad Sciences)

Ulloa, B. F. L., Møller, V., & Sousa-Poza, A. (2013). How does subjective well-being evolve with age? A literature review. *Journal of Population Ageing*, *6*(3), 227–246. (Publisher: Springer)

Van Landeghem, B. (2012). A test for the convexity of human well-being over the life cycle: Longitudinal evidence from a 20-year panel. *Journal of Economic Behavior & Organization*, *81*(2), 571–582. (Publisher: Elsevier)

Yang, Y., & Land, K. C. (2008). Age–period–cohort analysis of repeated cross-section surveys: fixed or random effects? *Sociological methods & research*, *36*(3), 297–326. (Publisher: Sage Publications Sage CA: Los Angeles, CA)

Chapter 6 Appendix

Table A6.1: Survey questions for well-being and mental health measures.

Measure	Question
Life satisfaction	On a scale from 0 to 10 where 0 means completely dissatisfied and 10 means completely satisfied, how satisfied are you with your life?
CASP-12[a]	How often, if at all, have you experienced the following feelings and thoughts over the past four weeks:
Control	How often do you think your age prevents you from doing the things you would like to do?
	How often do you feel that what happens to you is out of your control?
	How often do you feel left out of things?
Autonomy	How often do you think that you can do the things that you want to do?
	How often do you think that family responsibilities prevent you from doing what you want to do?
	How often do you think that shortage of money stops you from doing the things you want to do?
Pleasure	How often do you look forward to each day?
	How often do you feel that your life has meaning?
	How often, on balance, do you look back on your life with a sense of well-being?
Self-Realization	How often do you feel full of energy these days?
	How often do you feel that life is full of opportunities?
	How often do you feel that the future looks good for you?
EURO-D[b]	Earlier we talked about your physical health. Another measure of health is your emotional health or well-being that is, how you feel about things that happen around you.
Depression	In the last month, have you been sad or depressed?
Pessimism	What are your hopes for the future?
Suicidality	In the last month, have you felt that you would rather be dead?
Guilt	Do you tend to blame yourself or feel guilty about anything[c]?
Sleep	Have you had trouble sleeping recently?
Interest	In the last month, what is your interest in things[d]?
Irritability	Have you been irritable recently?
Appetite	What has your appetite been like[e]?
Fatigue	In the last month, have you had too little energy to do the things you wanted to do?
Concentration	How is your concentration? For example, can you concentrate on a television program, film or radio program?
	Can you concentrate on something you read?
Enjoyment	What have you enjoyed doing recently?
Tearfulness	n the last month, have you cried at all?

[a]Index generated from questions on 4 different dimensions. The total score ranges from 12 (low quality of life) to 48 (high quality of life). The response options for each item are: 1. Often, 2. Sometimes, 3. Rarely, and 4. Never.
[b]Index generated from questions on 12 different dimensions. The total score ranges from 0 (not depressed) to 12 (very depressed). The responses are coded as: 0. No indication and 1. There is indication of the respective dimension.
[c]If the answer is unclear the follow-up question is: So, for what do you blame yourself?
[d]If the answer is unclear the follow-up question is: So, do you keep up your interests?
[e]If the answer is unclear the follow-up question is: So, have you been eating more or less than usual?

Figure A6.1: Coefficients of the age dummies model (6.1) for the different countries.

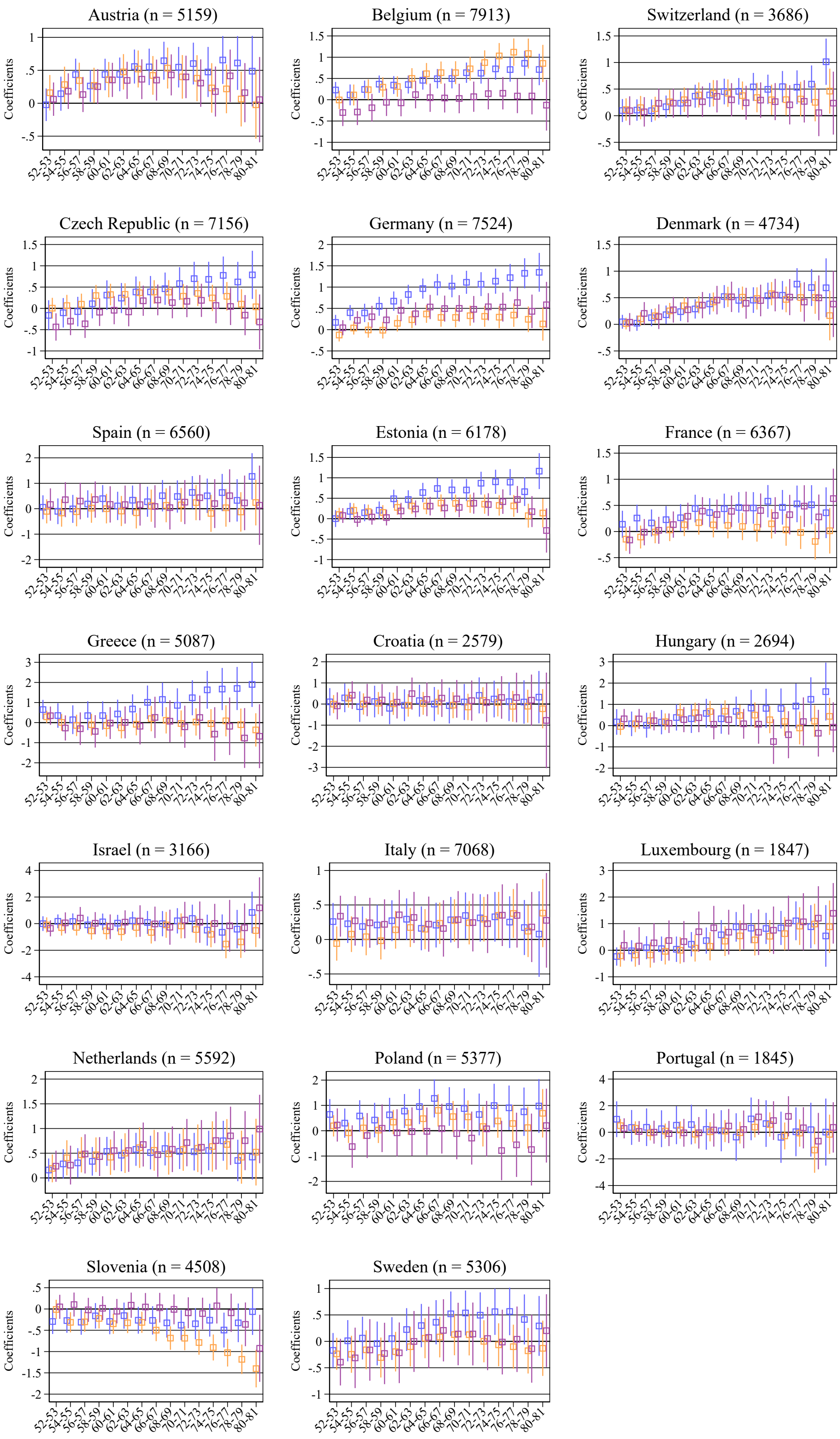


Table A6.2: Coefficients of the pooled OLS model (6.2) (using year dummies).

		Unweighted			Using sampling weights		
		No controls	Ex. controls	All controls	No controls	Ex. controls	All controls
Life satisfaction	Age	0.173^{***}	0.131^{***}	0.0659^{***}	0.146^{***}	0.115^{***}	0.0684^{**}
		(17.28)	(13.61)	(6.34)	(6.57)	(5.41)	(3.12)
	Age^2	-0.00136^{***}	-0.00102^{***}	-0.000353^{***}	-0.00123^{***}	-0.000928^{***}	-0.000402^{*}
		(-17.39)	(-13.74)	(-4.38)	(-7.10)	(-5.63)	(-2.36)
CASP-12	Age	0.242^{***}	0.213^{***}	0.213^{***}	0.250^{***}	0.205^{***}	0.161^{***}
		(25.45)	(24.32)	(16.15)	(12.41)	(10.53)	(8.50)
	Age^2	-0.00204^{***}	-0.00180^{***}	-0.00115^{***}	-0.00218^{***}	-0.00181^{***}	-0.00127^{***}
		(-27.55)	(-26.47)	(-15.60)	(-13.99)	(-11.99)	(-8.66)
EURO-D	Age	0.274^{***}	0.233^{***}	0.154^{***}	0.272^{***}	0.210^{***}	0.149^{***}
		(26.59)	(23.25)	(13.37)	(11.58)	(9.42)	(6.51)
	Age^2	-0.00225^{***}	-0.00194^{***}	-0.00113^{***}	-0.00227^{***}	-0.00178^{***}	-0.00112^{***}
		(-27.77)	(-24.68)	(-12.64)	(-12.35)	(-10.24)	(-6.27)

Notes: t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A6.3: Correlation between happiness measures and death, dependent variable is probability of dying between waves, logit model (6.4).

	Life satisfaction	CASP-12	EURO-D
Life satisfaction	-0.0576 (-1.80)		
CASP-12		-0.0768* (-2.04)	
EURO-D			-0.0773** (-2.58)
Marriage/registered partnership	-0.0126 (-0.10)	-0.0205 (-0.15)	-0.0359 (-0.29)
Income brackets:			
[1] Average monthly income per hh, low to mid bracket	0.392** (2.59)	0.411** (2.62)	0.557*** (3.82)
[2] Average monthly income per hh, mid to high bracket	0.357* (2.24)	0.386* (2.35)	0.488** (3.22)
[3] Average monthly income per hh, more than high bracket	-0.322 (-1.81)	-0.291 (-1.57)	-0.274 (-1.62)
Education:			
[1] Primary school	-0.257 (-0.90)	-0.405 (-1.40)	-0.411 (-1.52)
[2] Lower secondary school	-0.472 (-1.65)	-0.550 (-1.92)	-0.483 (-1.79)
[3] Upper secondary school	-0.784** (-2.76)	-0.842** (-2.95)	-0.767** (-2.87)
[4] Post-secondary non-tertiary education	-0.366 (-1.00)	-0.499 (-1.34)	-0.377 (-1.09)
[5] First stage tertiary education	-0.670* (-2.20)	-0.708* (-2.31)	-0.675* (-2.34)
[6] Second stage tertiary education	-1.429 (-1.26)	-1.454 (-1.28)	-1.413 (-1.28)
Subjective health:			
[1] Fair	-1.498*** (-9.99)	-1.505*** (-9.83)	-1.465*** (-10.19)
[2] Good	-2.431*** (-13.70)	-2.436*** (-13.09)	-2.378*** (-13.79)
[3] Very good	-2.653*** (-10.27)	-2.745*** (-10.05)	-2.628*** (-10.37)
[4] Excellent	-2.341*** (-7.42)	-2.296*** (-7.01)	-2.343*** (-7.57)
Constant	-6.300*** (-12.92)	-6.199*** (-12.51)	-5.364*** (-12.74)
N	164134	160305	131920

Notes: t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A6.4: Coefficients of the pooled OLS model (6.2) with interaction terms for the in all waves dummy.

	Life satisfaction	CASP-12	EURO-D
In all waves	-1.700	-4.015*	-0.0567
	(-0.94)	(-2.17)	(-0.03)
Age	0.0885***	0.149***	0.131***
	(6.52)	(12.06)	(8.14)
In all waves # Age	0.0536	0.116*	0.00859
	(1.00)	(2.11)	(0.14)
Age^2	-0.000502***	-0.00110***	-0.000959***
	(-4.72)	(-11.33)	(-7.57)
In all waves # Age^2	-0.000409	-0.000808*	-0.0000980
	(-1.04)	(-2.00)	(-0.22)
N	164125	160296	131913
R^2	0.236	0.349	0.253

Notes: t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A6.5: Coefficients of the pooled OLS model (6.2) with interaction terms for the female dummy.

	Life satisfaction	CASP-12	EURO-D
Female	2.140	0.659	0.574
	(1.53)	(0.55)	(0.39)
Age	0.140***	0.166***	0.137***
	(3.85)	(5.16)	(3.55)
Female # Age	-0.0664	-0.0209	-0.0386
	(-1.51)	(-0.55)	(-0.84)
Age^2	-0.000890**	-0.00126***	-0.00101***
	(-3.13)	(-5.01)	(-3.32)
Female # Age^2	0.000528	0.000162	0.000326
	(1.54)	(0.56)	(0.92)
N	163703	159880	131523
R^2	0.236	0.324	0.251

Notes: t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A6.6: Coefficients of the pooled OLS model (6.2) for different countries.

		Life satisfaction	CASP-12	EURO-D
AUT (N = 5159) Austria	Age	0.206**	0.268***	0.192**
		(3.19)	(4.27)	(3.12)
	Age^2	-0.00143***	-0.00207***	-0.00142**
		(-2.80)	(-4.17)	(-2.91)
BEL (N = 7913) Belgium	Age	0.0516	0.056	0.0241
		(0.94)	(0.88)	(0.28)
	Age^2	-0.000190	-0.000133	-0.000104
		(-0.45)	(-0.27)	(-0.16)
CHE (N = 3686) Switzerland	Age	0.0440	0.168***	0.173**
		(0.95)	(3.51)	(3.02)
	Age^2	-0.000177	-0.00125***	-0.00131**
		(-0.49)	(-3.37)	(-2.92)
CZE (N = 7156) Chech Republic	Age	0.128*	0.256***	0.234***
		(2.05)	(5.29)	(3.64)
	Age^2	-0.000691	-0.00191***	-0.00168***
		(-1.44)	(-5.15)	(-3.41)
DEU (N = 7524) Germany	Age	0.263***	0.166***	0.180**
		(4.92)	(3.51)	(2.90)
	Age^2	-0.00170***	-0.00118**	-0.00126**
		(-4.08)	(-3.20)	(-2.64)
DEN (N = 4734) Denmark	Age	0.0682	0.194***	0.135*
		(1.40)	(4.35)	(2.16)
	Age^2	-0.000353	-0.00135***	-0.000932
		(-0.92)	(-3.82)	(-1.92)
ESP (N = 6560) Spain	Age	0.0552	0.0453	-0.0205
		(0.55)	(0.43	(-0.14)
	Age^2	-0.000140	-0.000254	0.000163
		(-0.18)	(-0.31)	(0.14)
EST (N = 6178) Estonia	Age	0.157**	0.164***	0.122*
		(3.24)	(4.06)	(2.35)
	Age^2	-0.000948*	-0.00123***	-0.000868*
		(-2.50)	(-3.88)	(-2.11)
FRA (N = 6367) France	Age	0.0993	0.185***	0.144**
		(1.77)	(3.58)	(2.18)
	Age^2	-0.000654	-0.00142***	-0.000964
		(-1.51)	(-3.51)	(-1.86)
GRC (N = 5087) Greece	Age	0.00368	0.0433	0.293
		(0.03)	(0.39)	(1.58)
	Age^2	0.000357	-0.000442	-0.00225
		(0.31)	(-0.51)	(-1.55)

Notes: t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A6.6 (continued): Coefficients of the pooled OLS model (6.2) for different countries (cont.).

		Life satisfaction	CASP-12	EURO-D
HRV (N = 2579) Croatia	Age	-0.129	-0.0433	0.187
		(-0.96)	(-0.39)	(1.14)
	Age^2	0.00113	0.000301	-0.00146
		(1.09)	(0.35)	(-1.13)
HUN (N = 2694) Hungary	Age	-0.0836	0.298^{**}	0.0238
		(-0.63)	(3.07)	(0.20)
	Age^2	0.000932	-0.00229^{**}	-0.000283
		(0.89)	(-3.00)	(-0.30)
ISR (N = 3166) Israel	Age	-0.0994	-0.0775	-0.0955
		(-0.48)	(-0.44)	(-0.32)
	Age^2	0.000768	0.000330	0.000657
		(0.46)	(0.24)	(0.27)
ITA (N = 7068) Italy	Age	0.0650	0.0545	0.0476
		(1.06)	(0.95)	(0.64)
	Age^2	-0.000449	-0.000321	-0.000338
		(-0.96)	(-0.73)	(-0.60)
LUX (N = 1847) Luxembourg	Age	-0.0343	-0.0675	0.211
		(-0.29)	(-0.62)	(1.35)
	Age^2	0.000628	0.000855	-0.00143
		(0.66)	(1.01)	(-1.15)
NLD (N = 5592) The Netherlands	Age	0.208^{***}	0.183^{*}	0.113
		(3.54)	(2.45)	(1.39)
	Age^2	-0.00149^{***}	-0.00135^{*}	-0.000696
		(-3.34)	(-2.33)	(-1.11)
POL (N = 5377) Poland	Age	0.372^{**}	0.338^{**}	0.453^{*}
		(2.70)	(2.71)	(2.20)
	Age^2	-0.00268^{*}	-0.00250^{**}	-0.00354^{*}
		(-2.52)	(-2.59)	(-2.20)
POL (N = 5377) Poland	Age	0.0491	0.128	0.0951
		(0.17)	(0.68)	(0.37)
	Age^2	-0.000898	-0.00122	-0.000736
		(-0.40)	(-0.80)	(-0.35)
PRT (N = 1854) Portugal	Age	-0.0186	0.140^{**}	0.0876
		(-0.30)	(2.74)	(1.24)
	Age^2	0.000122	-0.00136^{***}	-0.000786
		(0.25)	(-3.44)	(-1.41)
SWE (N = 5306) Sweden	Age	0.227^{***}	0.254^{***}	0.237^{**}
		(3.50)	(4.21)	(3.12)
	Age^2	-0.00149^{**}	-0.00186^{***}	-0.00172^{**}
		(-3.07)	(-4.13)	(-3.02)

Notes: t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

List of Figures

List of Tables