Essays in Applied Microeconomics

Inauguraldissertation

zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften

durch die

Rechts- und Staatswissenschaftliche Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Axel Wogrolly

aus Graz, Österreich

2021

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Hans-Martin von Gaudecker
Zweitreferent:	Prof. Dr. Thomas Dohmen
Tag der mündlichen Prüfung:	18. Dezember 2020

Acknowledgements

Writing this dissertation has been a formidable challenge for me. Doubts that I could overcome it were a constant companion throughout my time as a doctoral student. That the endeavour did not end in failure is, in no small part, thanks to the support I've had every step of the way.

I want to thank my co-authors Hans-Martin von Gaudecker and Christian Zimpelmann. It has been hugely enjoyable, and intellectually rewarding, to do research with them. I'm especially grateful to Hans-Martin, my first supervisor, for his support and for helping me bridge the gap between being a reader of research and a researcher. I'm also indebted to Thomas Dohmen, my second supervisor, for invaluable feedback and guidance on my third project, and to Dominik Liebl for helpful advice and for agreeing to serve as the chair of my dissertation committee.

Without the tireless work of the administrative staff at the BGSE and the Institute of Applied Microeconomics, these places would not have been as conducive to doing research as I experienced them. I want to acknowledge Britta Altenburg, Simone Jost, Silke Kinzig and Andrea Reykers in this regard.

I'll look back at the period I spent alongside my colleagues at the BGSE as one characterised by friendship and camaraderie, not rivalry or competition. I'll remember many pleasant conversations over lunch or drinks I shared with Si Chen, Carl Heese, Marek Ignaszak, Marta Kozakiewicz, Peng Liu, Thomas Neuber and Gašper Ploj.

Most of all, I want to thank my parents, Albert and Brigitte, and my brother, Jakob, for always having my back.

Contents

List of Figures

List of Tables

Introduction

Many of the most important decisions in life are made from a place of uncertainty. Consider the question of how much of one's savings should be invested in the stock market. The answer hinges on how stock prices will evolve, which is unknown at the time the decision has to be taken and subject to many possible future influences, ranging from technological breakthroughs to pandemics. How do individuals deal with such uncertainty? How should they? This thesis consists of three self-contained essays that aim to shed light on these questions.

The benchmark model of economics, expected utility theory, assumes that individuals manage uncertainty by forming beliefs about all relevant events and that these beliefs satisfy the probability axioms. Knight (2012) observed that whilst the risk of coinflips is readily quantified, the uncertainty of natural events is not. There is thus ample scope for individuals to come to radically different conclusions on the likelihood of events. The first chapter of this thesis examines differences between individuals in how their beliefs evolve in the context of stock returns. In this chapter, which is joint with Hans-Martin von Gaudecker, we analyse a long panel of households' stock market beliefs to gain insights into the nature of the levels, dynamics, and informativeness of these expectations. In a first step, we classify respondents into one of five groups based on their beliefs data alone. In a second step, we estimate models of expectations at the group level so that belief levels, volatility, and response to information can vary freely across groups. At opposite extremes in terms of optimism, we identify pessimists who expect substantially negative returns and financially sophisticated individuals whose expectations are close to the historical average. Two groups expect average returns around zero and differ only in how they respond to information: Extrapolators who become more optimistic following positive information and mean-reverters for whom the opposite is the case. The final group is characterised by its members being unable or unwilling to quantify their beliefs about future returns.

Since Ellsberg (1961), we have known that expected utility theory cannot fully account for how individuals choose under uncertainty. In some settings, observed decisions of a sizable fraction of individuals cannot be explained with any probabilistic belief. This suggests that heterogeneity in how individuals manage uncertainty goes beyond beliefs, extending to what has become known as ambiguity attitudes. Analysis of these attitudes for natural events is the subject of the second chapter, which is joint with Hans-Martin von Gaudecker and Christian Zimpelmann. We analyse the stability and distribution of ambiguity attitudes using a representative sample. We employ four waves of data from a survey instrument with high-powered incentives. Structural estimation of random utility models yields three individual-level parameters: Ambiguity aversion, likelihood insensitivity or perceived level of ambiguity, and the variance of decision errors. We demonstrate that these parameters are very heterogeneous but fairly stable over time and across domains. These contexts span financial markets, our main application, and climate change. The ambiguity parameters are interdependent in their interpretation and the precision of their estimates depends on decision errors. To describe heterogeneity in these three dimensions, we adopt a discrete classification approach. A third of our sample comes rather close to the behaviour of expected utility maximisers. Half of the sample is characterized by a high likelihood insensitivity, with thirty per cent ambiguity-averse and twenty per cent making ambiguity-seeking choices for most events. For the remaining eighteen per cent, we estimate sizeable error parameters, which implies that no robust conclusions about their ambiguity attitudes are possible. Predicting group membership with a large number of observed characteristics shows reasonable patterns.

The difficulty of finding reliable probabilities for uncertain events raises the question how individuals aspiring to be rational should approach it. One intriguing proposal is to turn to prediction markets (Arrow, Forsythe, Gorham, Hahn, Hanson, et al., 2008) or bookmakers. In the third chapter, I examine betting odds, which are often seen as a credible source of predictions for future events in sports, politics, and entertainment. Who will win Wimbledon, who will be the next US President and which movie will win Best Picture at the Oscars? Betting odds, offered by bookmakers or by traders in prediction markets, typically exist for answers to each of these questions and can be turned into implied probabilities. Are these well-calibrated in the sense that they indicate the empirical frequency of outcomes? Do they incorporate publicly available information that might be relevant for prediction? Using a large sample of ATP tennis matches, I investigate these questions with machine learning methods that combine model-based ratings of player strength with a large number of other player features to estimate probabilities. I find that, in almost all settings, implied probabilities are very well calibrated and can be regarded as probabilities of events that condition on an information set containing publicly available statistics on players and matches. They reduce the error of the best prediction model by around 1.3% in terms of negative log-likelihood.

References

Arrow, Kenneth J, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D Nelson, et al. 2008. "The promise of prediction markets." *Science* 320: 877–78. [2]

Ellsberg, Daniel. 1961. "Risk, Ambiguity, and the Savage Axioms." *The Quarterly Journal of Economics* 75 (4): 643–69. [1]

Knight, Frank H. 2012. *Risk, uncertainty and profit*. Courier Corporation. [1]

Chapter 1

Heterogeneity in households’ stock market beliefs: Levels, dynamics and epistemic uncertainty[*]

Joint with Hans-Martin von Gaudecker

1.1 Introduction

Understanding households’ stock market expectations is critical for models of life-cycle behaviour, portfolio choice, and asset pricing. A number of key facts have been established for the cross-section of subjective beliefs about equity returns (Manski, 2004; and Hurd, 2009, provide excellent overviews, we pay detailed credit below). Beliefs differ widely across individuals. On average, they tend to be pessimistic relative to historical returns. Stated beliefs exhibit focal point responses; when it comes to probabilities, 50:50 is a particularly common answer. Stated expectations of a sizeable fraction of individuals are not consistent with the laws of probability. Optimism and consistency of beliefs are positively related to socio-economic variables in general and measures of financial sophistication in particular.

More recently, additional attention has been paid to the process of belief formation as a potential source of this heterogeneity. Taking a long-term perspective, Malmendier and Nagel (2011) show that individuals who experienced larger stock returns over the course of their lives tend to expect larger future returns. Greenwood and Shleifer

[*] We would like to thank Stéphane Bonhomme, Andreas Dzemski, Charles Manski, Simas Kucinskas, and Joachim Winter as well as participants of the first CRC-TR 224 conference, the Ifo/LMU/CRC-TR 190 Workshop on Subjective Expectations and Probabilities in Economics, and the 2018 Econometric Society European Meetings, for their comments. We thank Matthias Vollbracht from Media Tenor International for giving us access to data and Andreas Dzemski for sharing and explaining the code calculating the test statistic from Dzemski and Okui (2018). Both authors are grateful for financial support by the German Research Foundation (DFG) through CRC-TR 224 (Project C01)

(2014) find that on average, beliefs extrapolate recent stock market performance into the future. Adam, Marcet, and Beutel (2017) test the rational expectations hypothesis using subjective expectations data and reject it. Barberis, Greenwood, Jin, and Shleifer (2015) and Adam, Marcet, and Nicolini (2016) develop asset pricing models that feature investors with non-standard belief formation processes, showing that this matters for aggregate outcomes.

Starting from these sets of observations, this paper estimates processes for the formation of households' stock market beliefs, taking into account heterogeneity in levels, volatility, response to information, and epistemic uncertainty. We make use of an unusually long panel of probabilistic belief statements in the RAND American Life Panel, which was commissioned by and first described in Hurd and Rohwedder (2011). We start by verifying in our data the key facts in the cross section and on average belief formation, expanding upon them in several directions. Most importantly, we add the tone of recent media reports on the economy in U.S. television as an additional source of information. We do so because respondents overwhelmingly cite the state of the economy as a driver of their return expectations while at the same time, many claim to not follow the stock market and report incorrect values for realised returns, making it unlikely that the behaviour of stock prices is their prime source of information.

Our analysis of belief heterogeneity focuses on four dimensions: Levels, volatility, response to recent stock market returns and economic news, and epistemic uncertainty. The average time series dimension of our data is 26, which is too short for estimation at the individual level. In order to allow for heterogeneity along the four dimensions, we employ the discretisation approach proposed in Bonhomme, Lamadon, and Manresa (2017). In a first step, we use the k-means clustering algorithm to assign individuals to groups based on the dependent variable. We use a number of individual-level moments relating to levels and volatility, its covariances with recent stock market returns and economic news, and measures of consistency and self-stated information content. These variables thus capture the four dimensions of interest. The procedure yields groups that are similar in spirit to the types studied in Dominitz and Manski (2011), whose beliefs also differ in their levels, volatility and response to recent stock market returns.

We focus on five groups in our main specification. Using less groups mixes individuals with very different economic behaviours; adding more leads to relatively little additional insights at the expense of making the results harder to summarise. We show that our groups are stable across specifications; varying important features of the sample or of the classifying procedure changes little. Results of the diagnostic tests for group membership by Dzemski and Okui (2018) further corroborate our choice of groups and modelling strategy. All groups are reasonably large with sizes ranging between 13% and 26% of the sample.

In a second step, we estimate models relating respondents' beliefs about future stock prices to past returns of the Dow Jones and the tonality of economic news, allowing parameters to fully vary across groups. We find that one group consists of individuals whose expectations are close to the historical performance of the stock market and who respond slightly positively to recent returns and news about the economy. They have very low rates of inconsistencies. This behaviour is the closest we get to rational expectations; we thus label them "sophisticates". Correlating group membership with other observables, they stand out for having better knowledge of financial markets and the stock market in particular. At the other extreme of average expectations, we estimate one group with substantially negative return expectations and little reaction to returns or news. We label them "pessimists"; they have average values for inconsistencies in the belief elicitation procedure. The latter also is true for two more groups who both have return expectations around zero. Of all groups, these two react the strongest to both returns and news, but in completely different ways. One expects recent trends to continue ("extrapolators"); the other expects them to revert again ("mean reverters"). The last group stands out from the rest in that its members frequently give 50:50 answers when asked about probability judgements and state that these are their way of expressing epistemic uncertainty in a follow-up question; their belief measures often violate the laws of probability calculus. We label this group "ignorants"; correlations with other characteristics reveal that its members indeed do not pay much attention to the stock market.

We show that the groups we identify have very different levels of stockholding and trading behaviour. The level of heterogeneity in trading profiles over our sample period arises because of our classification into groups based on (time-series) features of the dependent variable. Our findings are robust to a number of choices regarding the treatment of the data and to parameters of the classification procedure. Our approach of first grouping indivduals based on the dependent variable and estimating group-level models achieves much higher goodness of fit than using observables alone in a classical regression analysis. This is consistent with recent evidence from a mixed survey-administrative dataset in Giglio, Maggiori, Stroebel, and Utkus (2019), who document persistent heterogeneity in the levels of beliefs that is difficult to explain with observable characteristics.

In a final step, we use the method of Coibion and Gorodnichenko (2015) to test whether the expectations of any of our groups could be characterised as rational in the sense that their forecast errors are unpredictable. We find that this is not the case; all overreact to current information. This is in line with Bordalo, Gennaioli, Ma, and Shleifer (2018) who find evidence of overreaction for a range of macroeconomic variables, and unsurprising in light of how difficult it is to predict stock return better than the historical average does.

The rest of the paper is organised as follows. Section 2 describes our data, connects it to prior literature, establishes the key stylised facts for our data, and outlines our empirical strategy. In section 3, we present the results, including the descriptions of several robustness analyses, the details of which are relegated to the appendix. Section 4 concludes.

1.2 Data, stylised facts, and empirical strategy

We analyse data from the RAND American Life Panel (ALP, see https://alpdata.rand.org) that were collected between 2008 and 2016. The ALP is a panel representative of the U.S. population whose members are regularly interviewed over the Internet. Households lacking internet access upon recruitment were provided laptops to limit selection bias. In addition to providing a large set of background characteristics from regular surveys, the ALP serves as a laboratory for researchers who are able to collect data at low costs. Hurd and Rohwedder (2011) describe the first waves of the data that include the measures of stock market beliefs forming the core of our study; these are part of a survey module developed to assess the effects of the financial crisis on household behaviour and well-being. Next to many background variables, we are able to link several other surveys containing data on financial numeracy and knowledge, probability numeracy, and portfolio choices. Table 1.A.1 in the appendix contains the exact references for all variables that we use.

Table 1.1 contains summary statistics of the covariates we use in our main specification. Throughout the paper, we apply the same sampling restrictions, namely observing at least 5 waves of stock market beliefs. The age structure of our sample skews somewhat older than the adult population. Compared with the 2010 Census, our sample includes more individuals aged between 50 and 65 and less under the age of 30. Women are slightly overrepresented, and individuals in our sample are substantially better educated. The fraction of individuals whose highest educational attainment is high school and below is less than half of what it was in the population in 2010.

Our data include answers to several questions that probe subjects' engagement with the stock market. We use a measure of whether subjects participated in the stock market beyond retirement accounts (such as an IRA, 401(k) and similar). They were also asked to self-assess the extent to which they follow and understand the stock market. Table 1.1 shows that the majority of the respondents in our sample has not engaged much with the stock market. Three quarters do not own stocks outside of their retirement accounts. Less than half of respondents claim they follow the market; only 40% consider themselves to have a good understanding of it. For a subset of respondents, we also have a measure that explicitly tests their knowledge of past returns. Individuals were first asked to select the sign of the return or indicate that they do not know, then the magnitude by choosing one of several bins. As the actual returns were between 7% and 16% when respondents answered the question, we count answers of $[0\%, 10\%]$ and $[10\%, 20\%]$ as correct. 42% of respondents fall into this category. 7% estimate a larger value, 31% choose the "don't know" option and twenty percent give a negative sign.

The ALP data contain a standard battery of questions measuring financial literacy, which is a key predictor of financial decision making (Lusardi and Mitchell (2014)). We use data from a wave that was in the field between March and September 2009.

Table 1.1. Descriptive Statistics - Individual characteristics

Variable	Observations	Mean	Std. dev.	$q_{0.1}$	$q_{0.5}$	$q_{0.9}$
Age: ≤ 30	3030	0.14				
Age: $\in (30, 50]$		0.33				
Age: $\in (50, 65]$		0.37				
Age: ≥ 65		0.16				
Female	3030	0.59				
Education: High school or less	3030	0.18				
Education: Some college		0.38				
Education: Bachelor degree		0.26				
Education: Advanced degree		0.18				
Owns stocks	3030	0.27				
Follows stock market	3010	0.46				
Understands stock market	3010	0.40				
Knowledge of returns: False Sign	2067	0.20				
Knowledge of returns: Don't Know		0.31				
Knowledge of returns: Magnitude too large		0.07				
Knowledge of returns: Correct		0.42				
Financial Numeracy	1564	0.82	0.22	0.52	0.86	1
Financial Knowledge	1564	0.78	0.24	0.46	0.87	1
Probability Numeracy	1940	0.67	0.2	0.4	0.7	0.89

Notes: The observations summarised in the table are restricted to individuals in our final sample. For dummy variables, only means are shown. Age is set to the within-person median across surveys. Education is set to the within-person mode across surveys. "Owns stocks" is the within-person mean of a dummy equalling 1 if respondents indicated that their liquid portfolio included stocks or mutual funds. This excludes stock holdings as part of an IRA, 401(k), Keogh or similar retirement accounts.

"Follows stock market" equals 1 if individuals indicate they follow markets "very closely" or "somewhat" and 0 if "not at all". "Understands stock market" equals 1 if individuals rate their understanding of stock markets to be "extremely good", "very good" or "somewhat good" and 0 if they chose "somewhat poor", "very poor" or "extremely poor". The categories of "Knowledge of returns" refer to whether respondents were able to recall the return of the Dow Jones over the past year. Financial numeracy and knowledge are the first principle component for correct answers, rescaled to lie between 0 and 1, for the two sets of questions in the financial literacy battery referred to as basic and sophisticated in (Lusardi and Mitchell, 2007) Probability numeracy is the fraction of correct answers to questions aimed at measuring probabilistic reasoning (Hudomiet, Hurd, and Rohwedder, 2018).

The battery consists of two sets of questions aimed at measuring financial numeracy (often called "basic financial literacy") and financial knowledge ("advanced financial literacy"), respectively (eg Lusardi and Mitchell (2007)). We extract the first principal component from each block of questions and scale each measure to have support between zero and 1. Both measures are left-skewed and have means of 0.82 and 0.78, respectively.

Finally, we use the probability numeracy battery developed in Hudomiet, Hurd, and Rohwedder (2018), who find that few people understand complex laws of probability but that most people have a basic understanding. We limit ourselves to a basic measure by using the fraction of correct answers across questions an individual answered. Table 1.1 shows that the average fraction of correct responses is 0.67 with a standard deviation of 0.20, implying substantial variation in probability numeracy.

1.2.1 Measures of stock market beliefs

The data on stock market beliefs stem from the survey module "Effects of the Financial Crisis" (Hurd and Rohwedder, 2011), which was fielded between late 2008 and early 2016 with a total of 61 waves. The first two waves were collected in November 2008 and March 2009. Starting in May 2009, data were collected monthly until April 2013. Afterwards, the surveys ran at a quarterly frequency until they ended in January 2016. As we are interested in belief formation, we restrict ourselves to individuals who responded at least five times to the belief measures. In total, we have on average 26 waves of data for 3030 individuals for a total of 77310 observations available. Figure 1.A.1 in the appendix shows the distribution of survey waves by individual.

The belief measures we analyse consist of three points on the subjective cumulative distribution function. Let p_t be the value of the Dow Jones Industrial Average at time t, and $R_{t\to t+12} := \frac{p_{t+12}-p_t}{p_t}$ the return on the Dow Jones in 12 months. We are very explicit about the notation when it comes to timing because questions about annual returns are asked at a monthly or quarterly frequency, which may lead to confusion otherwise. All time indices in this paper indicate months. For $\Pr(R_{t\to t+12} > 0)$ the question was:

We are interested in how well you think the economy will do in the future. By next year at this time, what are the chances that mutual fund shares invested in blue chip stocks like those in the Dow Jones Industrial Average will be worth more than they are today?

For $\Pr(R_{t\to t+12} > 0.2)$ the question was:

By next year at this time, what is the percent chance that mutual fund shares invested in blue-chip stocks like those in the Dow Jones Industrial Average will have increased in value by more than 20 percent compared to what they are worth today?

For $\Pr(R_{t\to t+12} \leq -0.2)$ the question was:

By next year at this time, what is the percent chance that mutual fund shares invested in blue-chip stocks like those in the Dow Jones Industrial Average will have fallen in value by more than 20 percent compared to what they are worth today?

From the three points on the cumulative distribution function, we construct an approximation of an individual's expected return to serve as our primary dependent variable. The approximation is as follows:

$$E[R_{t\to t+12}] = \sum_{j}^{4} E[R_{t\to t+12} | R_{t\to t+12} \in I_j] \cdot \Pr(R_{t\to t+12} \in I_j)$$

where the intervals I_j are $[-\infty, -0.2], [-0.2, 0], [0, 0.2]$ and $[0.2, \infty]$. The probabilities in these expressions are observed in the data. We set the conditional means they average to the midpoint of each interval. For the open intervals, we set the lower and upper bounds to the 1^{st} and $99^t h$ percentiles of the historical distribution of the Dow Jones' return (−0.32 and 0.43, respectively). Rather than dropping sets of observations that violate monotonicity of the cumulative distribution function (i.e., $\Pr(R_{t\to t+12} \leq -0.2) \leq \Pr(R_{t\to t+12} \leq 0) \leq \Pr(R_{t\to t+12} \leq 0.2))$, we restore weak monotonicity by setting its values at -0.2 and/or 0.2 to its value at 0. Such monotonicity violations are very common in this question format—for example, around 40% of responses both in the data of Hurd, Rooij, and Winter (2011) and in our own. We will return to inconsistencies in the next section.

In robustness checks, we avoid assumptions on monotonicity violations altogether by focusing on the probability of a positive return (e.g. Dominitz and Manski, 2007, also had even though more more measures available and discarde them presumably for such reasons). Table 1.2 shows summary statistics for within-person means of the different belief measures, i.e., the mean return and the three points on the cumulative distribution function. We first calculate means for each individual and then average across individuals, thereby weighting every sample participant equally regardless of the number of times she participated. The variation across the different points of the distribution function appears reasonable and all measures exhibit substantial variation across individuals.

Table 1.2. Individual belief measures averaged over time

	Mean	Std. dev.	$q_{0.1}$	$q_{0.5}$	$q_{0.9}$
$E[R_{t\to t+12}]$	0.5	5.8	-6.9	0.6	8.1
$\Pr(R_{t\to t+12}) > -0.2$	74.6	13.4	55.0	76.5	90.9
$\Pr(R_{t\to t+12}) > 0$	44.0	17.8	19.4	45.3	67.9
$\Pr(R_{t\to t+12}) > 0.2$	26.8	14.2	9.1	25.3	47.1

Notes: N = 3030. Units in percentage points.

1.2.2 Stylised facts

Our data on stock market beliefs has a number of distinct features that motivate our modelling choices below. Most of these characteristics are similar to those in other data; we briefly highlight them here and provide a full set of statistics in appendix 1.A.2. We do present the co-movements of beliefs with recent information in the main text because these are of particular interest and, when it comes to economic news, novel.

Similar to findings summarised in Hurd (2009), average beliefs are well below historical returns. For example, the mean of individuals' expected returns in our data is 0.5%, compared to a historical value of 7.3%. Beliefs do not only vary across individuals as shown in Table 1.2, but also within individuals over time. The magnitude of within-variation is similar to the magnitude of between-variation. Regression analyses controlling for many other factors show that beliefs of financially sophisticated and knowledgeable individuals are more optimistic. They also reveal that their beliefs are more likely to constitute actual probability judgements in two different senses.

First, Bruin, Fischhoff, Millstein, and Halpern-Felsher (2000) argue that 50% answers might indicate that individuals are epistemically uncertain about an event rather than expressing subjective beliefs of equal likelihoods. Following up on that observation, the questionnaires that we use confront respondents who gave an answer equal to 50% for $\Pr(R_{t\to t+12} \leq 0)$ with a follow up question. It asks them to clarify whether they mean that the Dow Jones is equally likely to rise as it is to fall, or whether they want to express that they are unsure what to do (also see Enke and Graeber, 2019). 53% of all answers when the follow up question was encountered turn out to be best characterised as expressing uncertainty that way. Second, if respondents are unsure about the behaviour of the Dow Jones index, they will be more likely to give sets of answers that violate monotonicity. Regressions reported in appendix 1.A.2 show that even after controlling for numerous other characteristics, measures of stock market following, financial numeracy, and financial knowledge are associated with substantially lower rates of monotonicity violations.

A recent literature has documented that average return expectations covary with recent stock market movements. Kezdi and Willis (2008) and Hurd (2009) noted this phenomenon early on. Greenwood and Shleifer (2014) find evidence for it across a variety of data sets; they also coined the term "extrapolative expectations". We corroborate this finding. In addition, we find evidence that individuals, on average, react to other types of information. In a small-scale ALP survey that overlaps with individuals in our main data, respondents were first asked about the probability of a stock market gain, much in the same way as the first question reproduced in Section 1.2.1.[1] After a short interlude of questions not of interest to us, they were asked to state what they most thought about when answering this question. Figure 1.1 shows the distribution of possible answers; the state of economy is by far the most common answer.

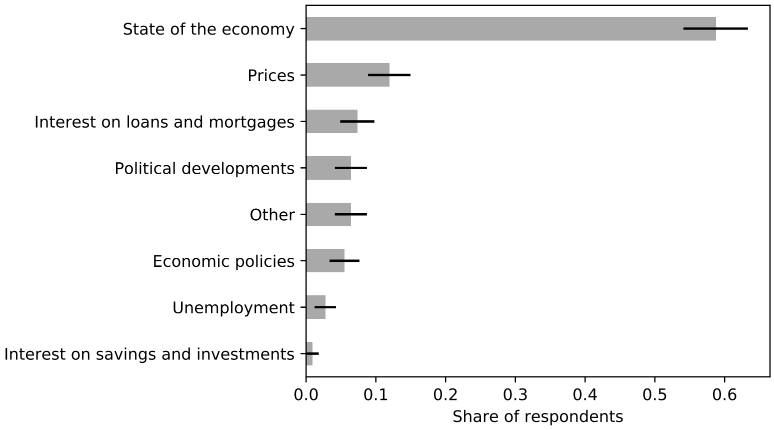


Figure 1.1. What respondents think most about when contemplating future stock prices

Notes: Only includes individuals who overlap with our main sample, $N = 114$.

This finding and the fact that only 42 percent of individuals in our sample have reasonable knowledge of how the Dow Jones changed over the preceding year (see Table 1.1) lead us to include additional information that subjects may use to form beliefs about stock returns. We hence obtained data on the tonality of economic news on major TV networks. We construct our measure using data provided by Media Tenor International, who had analysts classify evening news segments on CBS, Fox, and NBC in terms of what they refer to and whether the news is positive, neutral or negative. We take all news items referring to the state of the economy on day d and score positive items (pos) with 1, neutral items (neu) with 0 and negative times (neg) with - 1. We define our measure of the tonality of economic news as the average monthly score: $N_{t-1\rightarrow t} := \frac{\sum_{d\in[t-1,t]} 1\cdot\mathrm{pos}_d+0\cdot\mathrm{neu}_d-1\cdot\mathrm{neg}_d}{\sum_{d\in[t-1,t]} \mathrm{pos}_d+\mathrm{neu}_d+\mathrm{neg}_d}$.

We investigate the extent to which individuals extrapolate good and bad news, in form of recent stock returns and media reports on the economy as follows. We average expected returns across individuals for every survey wave, take first differences and plot them against the first differences of the Dow Jones Index return over the past month and the first differences of the average monthly news score. As shown in the first panel of Figure 1.2, an increase in the Dow Jones' returns over the past

[1] The precise question was "By next year at this time, what is the percent chance that mutual fund shares invested in blue chip stocks like those in the Dow Jones Industrial Average will be worth more than they are today?"

month of 4.3 percentage points (one standard deviation of the monthly return over our analysis period) is associated with a 0.13 percentage point higher expectation on the return over the next year. In the second panel depicting how first differences in expected returns vary with first differences in economic news, we see a similar pattern. A one unit increase in the news measure corresponds to an increase of 0.19 percentage points in expected returns. To put these numbers into context, Greenwood and Shleifer (2014) find that an increase in the annual return of 20 percentage points (one standard deviation of the annual return over the period on which their regression is based) increases the Michigan Survey expectations 0.78 percentage points [2].

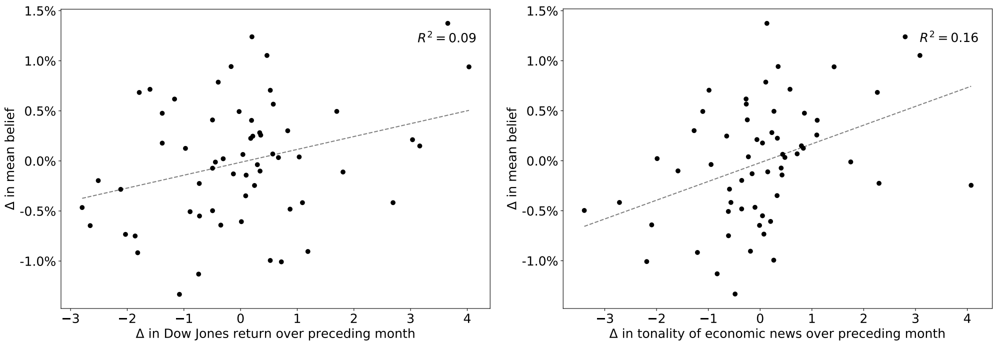


Figure 1.2. Average expected returns extrapolate stock prices and follow the tone of news

Notes: Both panels depict survey-to-survey changes in means of expected returns on the y-axis. The x-axis depict survey-to-survey changes in the standardised measures of recent monthly returns and news, respectively. X-axis units in time series standard deviations.

1.2.3 Empirical strategy

The stylised facts about individuals' stock market beliefs have shown that beliefs are very heterogeneous within and across individuals; that part of the between-variation is explained by financial sophistication; that the beliefs' evolution over time covaries with past returns and news about the economy; and that measures of beliefs vary in their informational content about true beliefs, which again varies systematically with financial sophistication. Together, these facts point towards putting between-person heterogeneity at the centre of a model of beliefs and their evolution over time. In particular, models that treat heterogeneity as an incidental parameters problem—fixed effects estimation being arguably the most prominent example—are doomed to fail. We expect individuals to differ in their levels of beliefs, in their belief volatility

[2] Greenwood and Shleifer (2014) obtain these results by regressing expected returns on annual returns. Figure 1.2 depicts the first differenced version of that regression, replacing annual with monthly returns, and separately also for news

over time, in how they change their beliefs in response to information about recent returns and economic news, and in the extent to which the measures we have at our disposal represent actual, accurate beliefs. At the same time, we need to impose some restrictions across individuals because our panel dimension is too short to allow for estimating models at the level of the individual.

We thus assume that we can summarise heterogeneity in belief formation processes by using a discrete set of groups. As long as the number of groups does not become too large, it allows us to describe the multidimensional patterns of heterogeneity in an accessible way; this would be difficult for many continuous distributions. Our main specification for belief formation is a linear model of the form:

$$E[R_{t\to t+12}]_{i,t} \;=\; \alpha_g + \sum_{l=0}^{L}\left(\beta_{g,l}R_{t-1-l\to t-l} + \gamma_{g,l}N_{t-1-l\to t-l}\right) + u_{i,t}. \tag{1.1}$$

We take $u_{i,t}$ to be independently and identically distributed across individuals and over time. We assume that all heterogeneity beyond that is captured by the coefficients. Put differently, we assume that there is a discrete number of groups G. All parameters of the model are allowed to differ at the group level, indexed by g: The intercept α_g measures the persistent degree of optimism or pessimism, the parameters $\beta_{g,l}$ measure how returns l months ago influence current beliefs, and $\gamma_{g,l}$ do the same for economic news N.

We estimate the model for $L = 0$, i.e., using only the most recent returns and news, and for $L = 6$. The latter allows us explore potential patterns of momentum in beliefs. We also experimented with averages across longer periods—e.g., much of the literature has considered annual returns—but found monthly intervals to provide the best fit. When constructing R and N, we are exact to the day on which individuals completed the survey.

In order to estimate the model, we employ the two-step method of Bonhomme, Lamadon, and Manresa (2017). In the first step, we classify individuals into a discrete set of G groups using moments of the both dependent and explanatory variables. In the second step, we estimate the coefficients in (1.1) separately for each group. This method is computationally simple and very transparent, providing easily interpretable groups.

Following Bonhomme, Lamadon, and Manresa (2017), we use the k-means algorithm in order to classify individuals into groups. The algorithm works by choosing the group assignments that minimise the sum of squared deviations between included variables and the group-wise means of these variables. The problem is NP-hard, but a number of heuristic algorithms exist that work well in practice. The method is widely used in machine learning; we use the implementation in the Python library *scikit-learn* (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, et al., 2011). Since solutions to the k-means objective are sensitive to the scaling of variables, we follow common practice and standardise each classification variable to have mean zero and unit variance in the cross-section of individuals.

In order to classify individuals into groups, we use moments of their stated beliefs and their relation with the explanatory variables. In particular, for each individual series of $\Pr(R_{t\to t+12} > -0.2)_{i,t}$, $\Pr(R_{t\to t+12} > 0)_{i,t}$, and $\Pr(R_{t\to t+12} > 0.2)_{i,t}$, we use its mean, its standard deviation, and its covariances with the return of the DJ as well as economic news, each measured over the month before the survey. These capture the dimensions level, volatility and response to information. In addition, we use the fraction of beliefs satisfying strict monotonicity, and the fraction of beliefs for which respondents did not indicate that beliefs expressed that they were unsure (or were not given the chance to do so). These capture the dimension of epistemic uncertainty. This makes for a total of fourteen time-constant moments that vary across individuals. We make this choice for two reasons. First, these moments exclusively use raw data and make no additional assumptions. This contrasts with, for example, expected returns, which entail a number of assumptions as detailed in Section 1.2.1. Second and more importantly, these are the key moments that should be informative on group-level heterogeneity along the dimensions we are interested in, as required for the analysis in Bonhomme, Lamadon, and Manresa (2017).[3]

Table 1.3. Moments and corresponding dimension

Moments	Dimensions
Mean probability that $R_{t,t\to t+12} \in (-0.2, \infty)$	Level
Mean probability that $R_{t,t\to t+12} \in (0, \infty)$	Level
Mean probability that $R_{t,t\to t+12} \in (0.2, \infty)$	Level
St. dev. of prob. that $R_{t,t\to t+12} \in (-0.2, \infty)$	Volatility
St. dev. of prob. that $R_{t,t\to t+12} \in (0, \infty)$	Volatility
St. dev. of prob. that $R_{t,t\to t+12} \in (0.2, \infty)$	Volatility
Cov. of prob. that $R_{t,t\to t+12} \in (-0.2, \infty)$ and returns	Response to Information
Cov. of prob. that $R_{t,t\to t+12} \in (0, \infty)$ and returns	Response to Information
Cov. of prob. that $R_{t,t\to t+12} \in (0.2, \infty)$ and returns	Response to Information
Cov. of prob. that $R_{t,t\to t+12} \in (-0.2, \infty)$ and news	Response to Information
Cov. of prob. that $R_{t,t\to t+12} \in (0, \infty)$ and news	Response to Information
Cov. of prob. that $R_{t,t\to t+12} \in (0.2, \infty)$ and news	Response to Information
Fraction of beliefs satisfying strict monotonicity	Epistemic Uncertainty
Fraction of beliefs expressing probability judgements	Epistemic Uncertainty

1.3 Results

We first outline describe the classification into groups, including a diagnostic test. We then describe our main results before reporting on a number of robustness checks.

[3] Note, however, the conceptual difference in that we assume that there is a discrete number of groups whereas the focus of the theoretical analysis in Bonhomme, Lamadon, and Manresa (2017) is on controlling for continuous unobservables.

The last part of this section explores the extent to which the groups we identify can be described by having rational expectations.

1.3.1 Classification into groups

In our main specification, we use five groups because this was the minimum number of groups where no economically meaningful intergroup differences were blurred. Larger numbers led to little additional economic insights and eventually to apparent overfitting. We will be more precise on this below in Sections 1.3.2 and particularly in 1.3.4, where we also consider alternative choices for the number of groups. As noted before, we require five belief measures per individual and use the moments listed in Table 1.3 as an input to the k-means algorithm.

Dzemski and Okui (2018) have developed a diagnostic test for clustering methods such as our classification step. Their procedure yields a unit-wise confidence set of group membership for each individual. It is constructed by testing the null hypothesis that individual i's true group g_i^0 is g for all groups $1, \dots, G$. The elements of the confidence sets are then those groups for which the null hypothesis cannot be rejected for a pre-specified confidence level. The test is based on the insight that if $g_i^0 = g$, then $E[(y_{i,t} - x_{i,t}^T \theta_g)^2)] \leq E[(y_{i,t} - x_{i,t}^T \theta_h)^2)]$ for all possible groups h (collecting all model parameters in the vector θ).

First of all, it is important to note that all group sizes are substantial. The largest group's share is 26% and that of the smallest is 13%. Figure 1.3 shows the distribution of unit-wise 90% confidence sets by their size and by whether they contain the estimated group. With 35% of individuals, the estimated group assignment being the only element in the set is the most common occurrence. For another twenty-three percent, the estimated group is in the confident set, but in addition to other groups. So for almost 60%, the estimated group is in the confident set. At the same time, very few confidence sets have more than three elements. Given that we have rather noisy data (compared to, say, the classification of states or countries, as the examples in Dzemski and Okui, 2018), these results demonstrate that our approach yields reasonable results even for a relatively low number of groups.

Nevertheless, a sizable fraction of confidence sets do not include the estimated group. Part of this is a reflection of the fact that the test is based on goodness of fit of our model (1.1), wheras the k-means procedure gives equal weight to all included features. Most notably, one would not expect the standard deviation over time to improve the fit of model (1.1). Indeed, the next section will demonstrate that level differences in expectations are the dominant component for improving model fit. Insofar as the k-means algorithm compromises splitting individuals along their expectation level to accommodate splitting them along differences in belief dispersion and association with returns or news, this will trivially result in a larger share being assigned to groups that are not in their confidence set than if one was assigning groups based on similar goodness-of-fit criteria as the test uses (as, for example, in

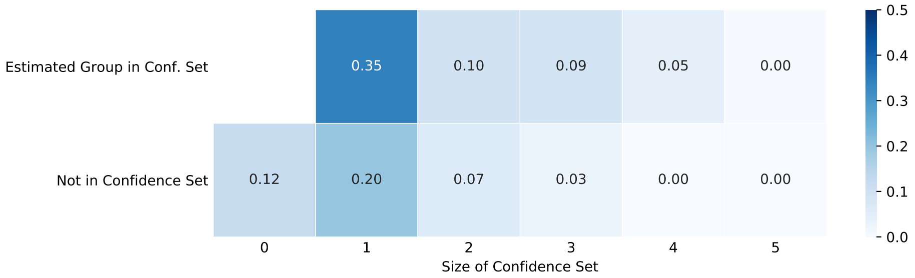


Figure 1.3. Unit-wise 10% confidence sets by size and inclusion of estimated group

Notes: The numbers refer to the share of individuals in each cell.

the clustering method of Bonhomme and Manresa, 2015). The reason confidence sets can be empty is similar; none of the estimated groups provide a good fit for some individuals whose expectation level is far from that of any of the groups.

1.3.2 Heterogeneity in groups' behaviour and characteristics

We order the 5 groups by their average expected returns and refer to them as pessimists, mean reverters, extrapolators, ignorants, and sophisticates, respectively. Each group's label captures the characteristic of the moments used for classification that makes it stand out from the others the most. The key results are summarised in three figures and two tables. Figure 1.4 plots the data averages (solid lines) versus the model predictions (dashed lines) of expected returns over time. Table 1.4 summarises the group means and standard deviations of expected returns, averaged within each group and survey wave, over our sample period. Figure 1.5 plots the reaction of groups to changes in past returns and news, respectively. Table 1.5 shows prevalence of monotonicity violations and the fraction of answers expressing epistemic uncertainty, respectively. Finally, Figure 1.7 presents the mean values of various covariates for each group.

Before describing each group in turn, we note that differences across all dimensions are important. The levels of beliefs in Figure 1.4 are strikingly different and—except for the mean reverters and extrapolators—hardly ever cross. The volatility over time is largest for mean reverters and extrapolators; it is by far smallest for ignorants with the other two groups in between. The reactions to both stock prices and news depicted in Figure 1.5 are substantial and very different.

Pessimists (25% of individuals) consistently expect the return of the Dow Jones to be negative and substantially so (-5.6%). Their beliefs do not vary too much over time, although they seem to be a bit more optimistic in the second half of our sample period (but still far below any other group). This seems to be due to better economic news in this period, to which they respond positively. Their beliefs do not react to

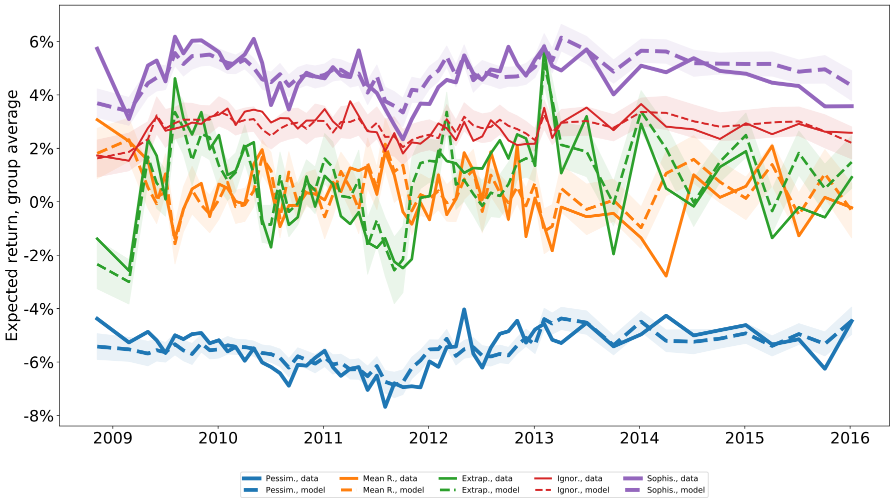


Figure 1.4. Data vs. predicted expected return of the Dow Jones index, by group

Notes: The solid and dashed lines are within survey and group means of individual data points and model predictions. Shaded regions are within survey × group means of individual 95% confidence intervals for the estimated regression function. Line widths are proportional to group sizes.

past return. Along the dimensions of knowledge and numeracy, pessimists appear to be in the middle of the distribution along with mean reverters and extrapolators.[4]

Mean Reverters and Extrapolators (19% and 17% of individuals, respectively) are also rather pessimistic, expecting a return of about zero. Individuals in these two groups are similar in observable characteristics. Their key difference, and reason for the labels we chose for them, can be seen in Figure 1.5: Extrapolators expect recent trends to continue and do so more than any other group. Mean reverters follow the opposite pattern: They become less optimistic following a good performance of the Dow Jones or positive economic news, and are the only group which reacts in this way. Hence, the lines in Figure 1.4 frequently cross and move in opposite directions survey to survey.

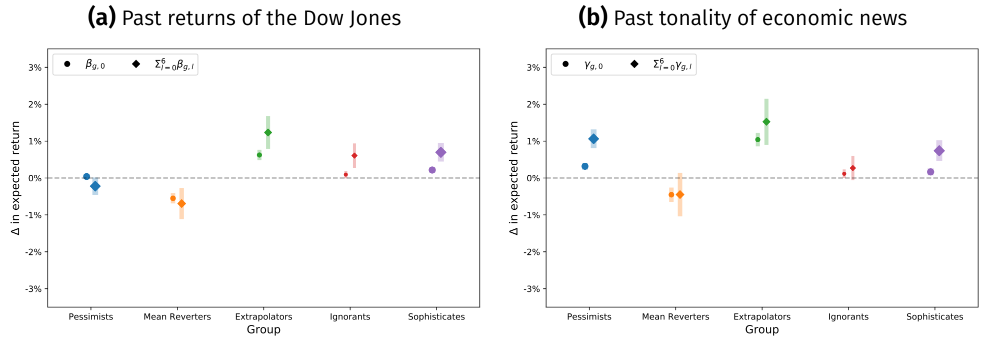
The fact that mean reverters and extrapolators are very similar in terms of observable characteristics, but react in completely different ways to information, under-

[4] The fractions of monotonicity errors and of beliefs expressing epistemic uncertainty (Table 1.5) seemingly stand in contrast to this, but they are probably due to somewhat mechanical effects. For monotonicity violations, giving low answers to $\Pr(R_{t\to t+12} > 0)$, as pessimist frequently do, will c.p. lead to less monotonicity errors if stated beliefs are subject to survey response error. This is because when stating the last elicited belief, $\Pr(R_{t\to t+12} < -0.2)$, the margin for avoiding a monotonicity error is larger when $\Pr(R_{t\to t+12} > 0)$ was small. In line with this explanation, the gap in monotonicity violations between pessimists and mean reverters / extrapolators is largely driven by violations of $\Pr(R_{t\to t+12} < -0.2) \leq Pr(R_{t\to t+12} < 0)$. In order to arrive at the follow-up question on epistemic uncertainty, an individual needs to use 50% when asked about the chance the Dow will increase. Pessimists feature very few 50% answers.

Table 1.4. Long run moments by group

	Data		Model	
	Mean	St. Dev.	Mean	St. Dev.
Pessimists	-5.57	0.80	-5.58	0.40
Mean Reverters	0.29	1.10	0.37	0.75
Extrapolators	0.79	1.81	0.91	1.43
Ignorants	2.78	0.49	2.80	0.15
Sophisticates	4.74	0.86	4.83	0.38

Notes: N = 3030. Units in percentage points. Expected returns are averaged within each group and survey wave, mean and standard deviation are calculated over the resulting time series points.


(a) Past returns of the Dow Jones

(b) Past tonality of economic news

Figure 1.5. Effect on expected returns of increases in past returns and tonality of economic news, by group

Notes: Dots depict the effect on expected returns of a one standard deviation increase in the most recent monthly return of the Dow Jones (Panel a) and in the most recent tonality of economic news over one month (Panel b). Diamonds depict the summed effect in the most recent, plus six preceding monthly returns of the Dow Jones and the preceding tonalities of economic news, respectively. Shaded lines show the width of 95% confidence intervals. Marker and line widths are proportional to group sizes.

lines the importance of classifying individuals in terms of features related to their stated beliefs. Considering only observed heterogeneity, as in classical regression analysis (see Section 1.C.1 of the appendix), would necessarily hide this important dimension of behaviour.

Ignorants (13% of individuals) are seemingly the second most optimistic group. Their average belief that the Dow Jones will increase is almost exactly 50% and they expect a return of 2.8%. Compared to the other groups, ignorants are notable for their very low belief variability. Panel B of Figure 1.7 shows that their average is near the tenth overall percentile and Figure 1.4 visualises how comparatively little individuals belonging to this group change their beliefs. As Figure 1.5 shows, their beliefs also covary least with returns and news. In addition to the belief that the Dow Jones will increase, the other two subjective beliefs are, on average, close to 50% as well which is incompatible with strong monotonicity of the cdf. Ignorants are most

Table 1.5. Measures of epistemic uncertainty by group

	Fraction of belief sets satisfying strict montonicity	Fraction of beliefs expressing subjective probabilities
Pessimists	0.67	0.96
Mean Reverters	0.42	0.90
Extrapolators	0.41	0.88
Ignorants	0.14	0.61
Sophisticates	0.75	0.95

Notes: N = 3030. A belief set satisfies strict monotonicity if $\Pr(R \leq -0.2) < \Pr(R \leq 0) < \Pr(R \leq 0.2)$. Beliefs express subjective probabilities if, for the question asking about the probability of an increase of the DJ, the belief is not 0.5, or it is and in the follow up question, the respondent indicated this means an equal likelihood.

likely out of all groups to violate monotonicity, with only 10% of belief sets satisfying it. One key reason for that is that where other groups express subjective probabilities with their stated beliefs 90% of the time and more, this is barely more than 60% for ignorants, which is below the tenth overall percentile. In other words, they use 50% answers to express epistemic uncertainty about stock returns. In line with their apparent lack of informedness, ignorants also have the lowest scores when it comes to following and understanding the stock market, knowledge of past returns and financial knowledge. Though seemingly more optimistic than other groups, all our indicators suggest that the stated beliefs of these individuals are limited in terms carrying quantitative information, and need to be interpreted with caution.

Sophisticates (26% of individuals), the most optimistic group, expect the Dow Jones to yield an average return of 5%. That number is relatively close to the historical performance of 7.3%. In addition to having beliefs that are most accurate compared to the historical distribution, sophisticates also stands out from the others in terms of experience with the stock market and knowledge relating to it. They are more likely to describe themselves as following and understanding the stock market, they have a superior knowledge of historical returns and greater financial knowledge. Sophisticates have the best understanding of probability calculus, are least likely to express beliefs that violate monotonicity of the cumulative distribution function (more than 80% of their belief sets satisfy strict monotonicity), and, together with pessimists, they use beliefs to express subjective probabilities most often.

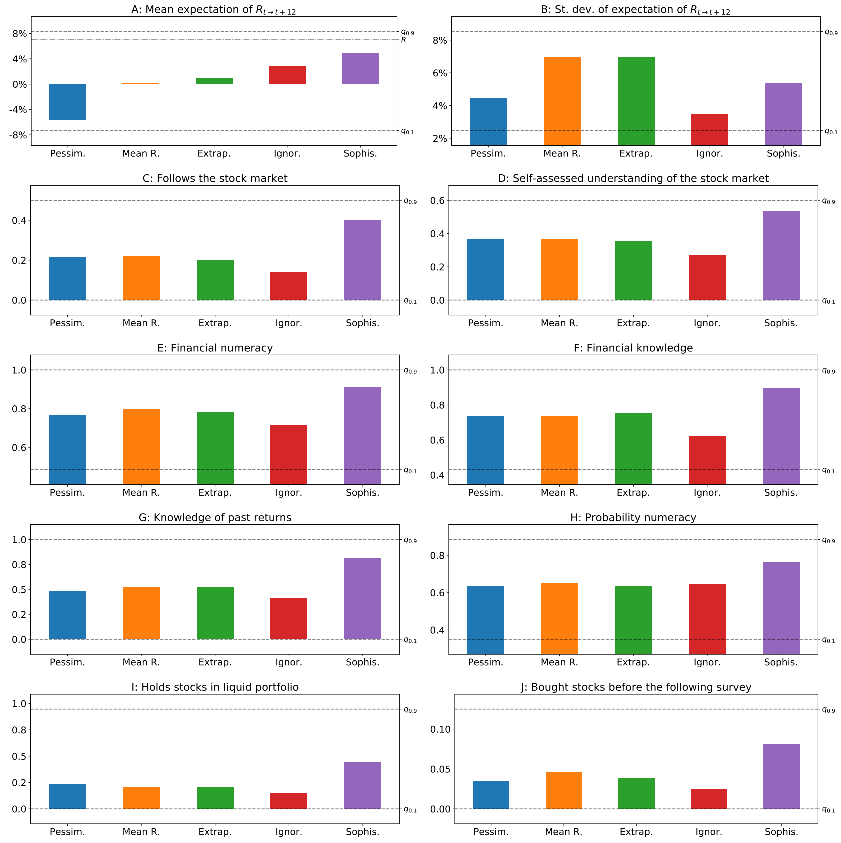


Figure 1.7. Observable characteristics by unobserved heterogeneity group

Notes: N = 3030, smaller for some panels depending on the availability of covariates, see Table 1.1. Bars show group means, dashed lines are the bottom and top decile with respect to the individuals of all groups taken together. *Variable definitions*:: Financial numeracy and knowledge: First principle components loading on variables indicating whether a respondent correctly answered numerical and knowledge based questions, scaled to the unit interval; Probability numeracy: Fraction of correct answers to questions about probability theory; Knowledge of past returns: False sign (0), don't know (⅓), magnitude too large (⅔), sign and magnitude correct (1); Understanding of the stock market: Extremely bad (0), very bad (⅕), bad (⅖), good (⅗), very good (⅘), extremely good (1); Follows stock market: Not at all (0), somewhat (½), closely (1).

1.3.3 Stock ownership and trading behaviour

The differences in beliefs and their trajectories translate into very different behaviour when it comes to portfolio choice. Panel I of Figure 1.7 shows that stockholding is lowest for ignorants (15%) and highest for sophisticates (44%), with the other groups below 20%, too. Trading behaviour follows a similar pattern (Panel J).

In order to further investigate this, we run a Probit regression of buying stocks in the subsequent period on a set of group fixed effects and the return expectation at the time of the survey. Because of the low baseline probabilities, it is important to use a nonlinear model as opposed to a linear probability model. This means that controlling for fixed effects is infeasible due to the incidental parameters problem. The average partial effects of increasing expectations by one group-level standard deviation are 0.14% for sophisticates and 0.03% for ignorants, respectively. For extrapolators, due to the comparatively greater volatility of their expectations (see Table 1.4) the effects are 0.16%. The other groups are somewhere in between.

Figure 1.8 shows that these patterns translate into very different predicted purchasing patterns over time. Pessimists and, even more so, ignorants hardly change their behaviour over time. Their predicted purchasing probabilities fluctuate slightly around low average values. The other three groups show much more differences over time. Again, this often goes in opposite directions. Not surprising, mean reverters would have higher than average values during the in the aftermath of the financial crisis, a time when the tone of news was also dire. Extrapolators show the opposite pattern. Sophisticates have the highest trading probability and a variability that is slightly below that of mean reverters or extrapolators.

It is once more important to note that these rich patterns of heterogeneous decision-making only surface because of our classification into groups based on (time-series) features of the dependent variable. Controlling for observed characteristics could only induce vertical shifts in trading behaviour, but no reversal of patterns. These are important, however, to generate potential trade between groups. Again, our findings mirror those reported in Giglio, Maggiori, Stroebel, and Utkus (2019), who find that beliefs to be reflected in portfolio allocations and a small but predictable effect of belief changes on trading patterns.

1.3.4 Discussion and robustness of results

Our preferred model explains more than a quarter of the variation in expected returns (see Table 1.B.2). This differs by a factor of one hundred from the same model without unobserved heterogeneity. Similarly, a standard regression model with lots of observed heterogeneity can explain only 12%.[5] This squares well with Giglio,

[5] See Table 1.C.1 for the precise results. The first regression is a linear probability model of the level of beliefs on the past six months' returns and news. The "kitchen-sink"-approach additional

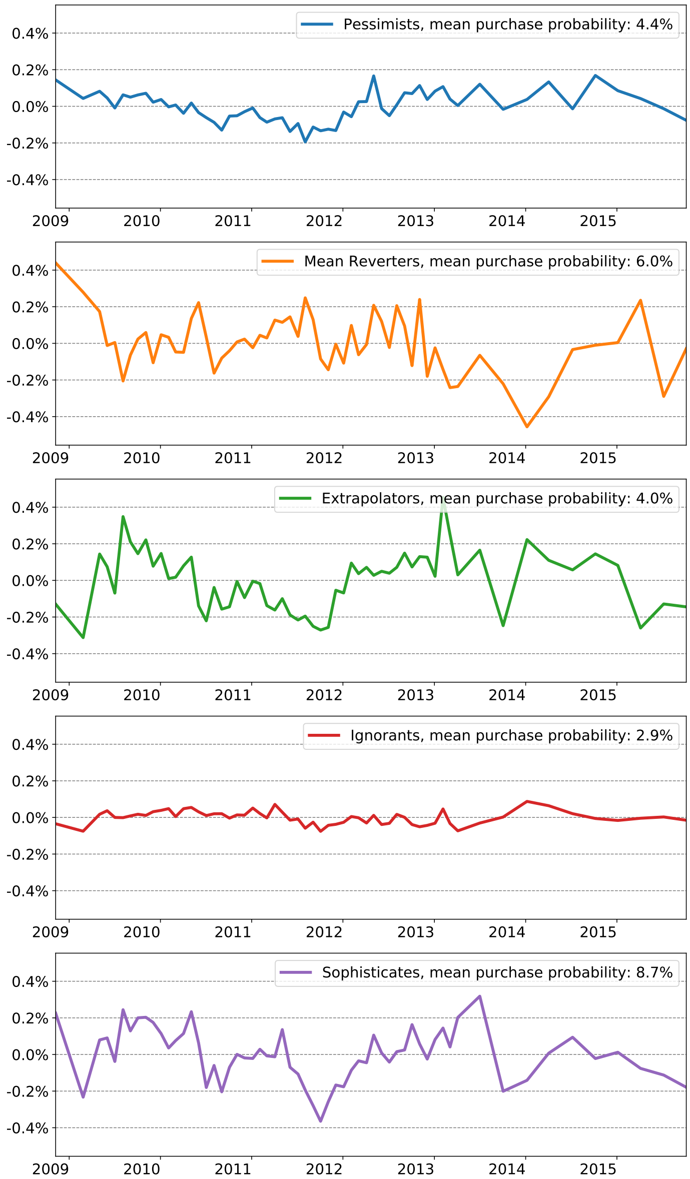


Figure 1.8. Predicted probability of buying stocks, group averages over time

Notes: N · T = 70114, N = 3029. Group and survey average predicted probabilities from a probit regression of a stock buying indicator in the next period, provided that is within 120 days, on group indicators and expected returns.

Maggiori, Stroebel, and Utkus (2019), who find as one of their five facts that "Beliefs are mostly characterized by large and persistent individual heterogeneity; demographic characteristics struggle to explain why some individuals are optimistic and some are pessimistic." Our analysis underlines their finding and goes beyond it in documenting different belief formation processes.

In Section 1.C.2 of the appendix, we relax the requirement of observing at least five sets of belief measures per individual to a minimum of three. The broad pattern of groups remains the same and we can essentially leave the labels in place (mean-reverters and extrapolators switch their places in terms of average expected returns and the latter group shrinks by about one third). Similarly, the group assignments remain very stable when requiring a minimum of fifteen periods per individual. Note that this ensures that the version of (1.1) with $L = 6$ lags would be identified individual-by-individual. The results are presented in Section 1.C.3 of the appendix. 86% of the respondents that meet the stricter requirement are assigned to the same group as before. The number is lowest for sophisticates at 73% (see Table 1.C.5), most of the remainder is assigned to the group of extrapolators.

As detailed in Section 1.2.1, our measure of expected returns makes a number of assumptions. In Section 1.C.4, we thus report results on a specification that uses the raw data on the probability of a stock market gain as the dependent variable. This also makes the analysis comparable to Dominitz and Manski (2011). By construction, the distribution of groups is exactly the same as in our main model (some tables and graphs shown in the other cases are thus superfluous) and, reassuringly, the diagnostic tests looks very similar, too. The time series look very similar to before with four clearly distinguishable levels; mean reverters and extrapolators are again on a similar level, crossing frequently. The reactions to simulated shocks show a similar pattern to Figure 1.5, if anything, lags seem to have slightly stronger effects towards building up momentum.

Sections 1.C.6, 1.C.7, 1.C.8, and 1.C.9 show the results for 3, 4, 7, and 15 groups, respectively. The results further motivate our choice of G. With three groups, the distinction between extrapolators and mean reverters is blurred with both mostly being allocated to the second group, resulting in extrapolation on average. In the case of four groups, the first four groups remain very stable (each retains more at least 94% of its previous members), but the group of sophisticates is distributed across the other groups with two thirds being pooled with pessimists (see Table 1.C.10). Based on those results, one may conclude that the most optimistic group was mostly made up by respondents with a severe lack of understanding or interest. It also blurs the features of the other groups; most notably, the average expectations of pessimists go up by two percentage points.

includes a quadratic in age, sex, education in four categories, ethnicity in five categories, 18 levels of household income, various measures of stock market experience and knowledge, probability and financial numeracy.

Moving from five to six or seven groups has effects almost exclusively for the groups of mean reverters, extrapolators, and ignorants. Both other groups retain more than three quarters of their members and all their characteristics remain very stable. Some of the clusters become fairly small and relative to the other lines in Figure 1.C.33, their data averages and predictions are very unstable. Consequently, the patterns become stronger, particularly on the extrapolation side. The positive interpretation of these patterns would be that some groups of individuals are reacting very strongly to current trends indeed; a sceptic may think that we are fitting noise in the data. In any case, we do not believe that one gains much additional insights from this relative to the case with five groups. The main reason for showing the results for fifteen groups is to demonstrate that while feasible, the algorithm clearly starts fitting noise. For example, group 12 consists only of 22 individuals. Note that the diagnostic test described in Section 1.3.1 becomes computationally infeasible.

1.3.5 Rational expectations tests

Greenwood and Shleifer (2014) and Giglio, Maggiori, Stroebel, and Utkus (2019) are just two examples of a large literature challenging the rational expectations paradigm for the average investor. In the light of our focus on heterogeneous belief formation processes, it seems very natural to ask whether some groups' belief formation processes may be consistent with rational expectations. In order to do so, we treat expectations as forecasts and analyse the predictability of forecast errors. We apply the methodology of Coibion and Gorodnichenko (2012, 2015), which yields a direct test of whether expectations are rational. In particular, forecast errors of full information rational expectations should be unpredictable with any information I_t at time t because they equal the true expected value of the variable to be forecasted given the information: $E\Big[R_{t\to t+12} - E[R_{t\to t+12} \,|\, I_t] \,|\, I_t\Big] = 0$. Non-full information rational expectation forecast errors should be unpredictable with any information in a forecaster's information set, though they might be with information the forecaster is not aware of or does not use. This insight allows for testing the rationality of expectations without knowing too much about either the true data generating process or what information forecasters use.

We follow the methodology of Coibion and Gorodnichenko (2015) who specify the information set I_t to be the forecast revision. Let $F_t R_{t\to t+h}$ be the forecast of the return $R_{t\to t+h}$ at time t of an individual. Forecast errors are then defined as $FE_t := R_{t\to t+h} - F_t R_{t\to t+h}$ and forecast revisions as $FR_t := F_t R_{t\to t+h} - F_{t-1} R_{t\to t+h}$. Regressing forecast errors on forecast revisions then tests the rationality of expectations, and the sign of the slope coefficient measures whether expectations overreact or underreact to information. If expectations are rational the slope coefficient is zero. A negative sign for the slope means an upwards revised forecast is typically followed by a downwards swing in the forecast error. As the regression includes an intercept, this means that the forecast overshoots, its upwards adjustments went too far. This is overreaction. The logic is reversed for a positive sign, which indicates underreaction.

To estimate this regression with our data, we have to make an assumption. Forecast revisions are defined as the difference of two forecasts of the return $R_{t\to t+12}$; this month's forecast $F_tR_{t\to t+12}$, for which we take individual expected returns, and last month's forecast $F_{t-1}R_{t\to t+12}$. We do not have a direct measure of the latter because beliefs were always elicited about one-year-ahead returns. To proceed, we assume that $F_{t-1}R_{t\to t+12} = F_{t-1}R_{t-1\to t-1+12}$. Hence we assume that last month's forecast of the return a year from then is also how respondents would have answered questions of the form: "What are the chances that mutual fund shares invested in blue chip stocks like those in the Dow Jones Industrial Average will be worth in thirteen months than what they will be worth in one month?". How strong is this assumption? Writing $R_{t-1\to t-1+12} - 1 = \frac{\frac{p_t}{p_{t-1}}}{\frac{p_{t+12}}{p_{t+11}}} \cdot (R_{t\to t+12} - 1)$, we see that it only depends on the next and last months. If individuals expect the same percentage change in stock prices over the next month as they do from 11 months ahead to 12 months ahead, the assumption is satisfied.

With this assumption, we can write the model as follows:

$$\begin{aligned} FE_{i,t} &= \tau_g + \delta_g FR_{i,t} + \epsilon_{i,t} \\ R_{t\to t+12} - E[R_{t\to t+12}]_{i,t} &= \tau_g + \delta_g \left(E[R_{t\to t+12}]_{i,t} - E[R_{t-1\to t+11}]_{i,t-1}\right) + \epsilon_{i,t} \end{aligned} \tag{1.2}$$

As before, we allow model coefficients to vary by group. Table 1.6 contains the results, restricting our sample to consecutive observations during the period where the survey was fielded monthly. Table 1.E.1 in the appendix repeats the exercise for our entire sample with very similar results.

As can be seen from the table, all groups overreact with a slope coefficient close to -0.5. This is exactly what we would find if time variation in expectations is uncorrelated with future returns[6]. Finding evidence of overreaction is unsurprising for two reasons. First, also using Coibion and Gorodnichenko (2015) regressions, Bordalo, Gennaioli, Ma, and Shleifer (2018) present evidence that overreaction of individual forecasters is prevalent across a wide range of macroeconomic variables. Second, stock returns are very difficult to predict. Campbell and Thompson (2007)

[6] Suppose forecasts and returns are uncorrelated and covariance stationary. Then δ equals exactly -0.5:

$$\begin{aligned} \delta &= \frac{cov(FE_t, FR_t)}{var(FR_t)} \\ &= \frac{cov(R_{t\to t+h} - F_tR_{t\to t+h}, F_tR_{t\to t+h} - F_{t-1}R_{t\to t+h})}{var(F_tR_{t\to t+h} - F_{t-1}R_{t\to t+h})} \\ &= -\frac{var(F_tR_{t\to t+h}) - cov(F_tR_{t\to t+h}, F_{t-1}R_{t\to t+h})}{2\cdot var(F_tR_{t\to t+h}) - 2\cdot cov(F_tR_{t\to t+h}, F_{t-1}R_{t\to t+h})} \;=\; -\frac{1}{2} \end{aligned}$$

Table 1.6. Predictability of forecast errors with forecast revisions

	Pooled OLS	Pooled OLS w groups
Forecast Revision	-0.52	
	(0.01)	
Forecast Revision, Pessimists		-0.52
		(0.01)
Forecast Revision, Mean Reverters		-0.51
		(0.02)
Forecast Revision, Extrapolators		-0.53
		(0.02)
Forecast Revision, Ignorants		-0.50
		(0.01)
Forecast Revision, Sophisticates		-0.53
		(0.02)
R^2	0.12	0.28
N · T	50532	50532

Notes: N = 2834. Observations that are not consecutive during the monthly phase of the survey waves are dropped. OLS estimates. Standard errors (clustered by individual and survey) in parentheses.

show predictive regressions fail to do better than the historical average unless augmented with theoretical restrictions. This points to a weak form of the efficient market hypothesis according to which one cannot use information to which typical U.S. citizens have access to form a forecast more accurate than forecasting the average return would be. In the previous section, we document that expectations react to recent returns and economic news on TV with sign and magnitude varying by group, and that they have sizable unexplained variation survey to survey on top of that. The results of Table 1.6 indicate that this variation in expectations is a form of overreaction.

1.4 Conclusions

We have analysed an unusually long panel of households' probabilistic stock market expectations collected in the RAND American Life Panel. Our first step was to document a number of key facts in these data, several of which have been known from other datasets and thus help establishing comparability. First, average beliefs are pessimistic relative to historical returns. Second, the dispersion of beliefs is very large, both across individuals in the cross-section and within individuals over time. Third, part of the variation over time is related to the fact that on average, beliefs extrapolate recent trends on the stock market. Fourth, individuals base their expectations for stock returns mostly on the state of the economy and the tone of recent media reports is positively related to average expectations. Fifth, the beliefs of financially sophisticated and knowledgeable individuals are more optimistic. Sixth, a non-trivial fraction of reported beliefs suffers from inconsistencies, part of which may be related to the fact that individuals truly have no quantitatively well-formed expectations. Finally, inconsistent beliefs are found less often for individuals who are financially sophisticated and knowledgeable.

Taking these facts as our point of departure, we have specified a simple model that relates beliefs to past returns and the tone of economic news. We have allowed for heterogeneity by first classifying individuals into one of five groups using the *k*-means clustering algorithm and then estimating the model separately for each group. The diagnostic test of Dzemski and Okui (2018) revealed that unit-wise confidence sets are small and that in 60% of the cases, they include the group we estimate individuals to be in. Only 12% behave in a way that is not captured by any of our groups, so that their confidence set is empty. This is despite the fact that our approach makes it difficult for the specification test in the sense that it is based on a very different statistic than what is used by the clustering algorithm.

Of our five groups, we have labelled the two polar cases in terms of optimism "pessimists" (annual return expectations well below zero, little reaction of expectations to either returns or news, average values for literacy indices) and "sophisticates" (annual return expectations close to the historical average, small positive reactions to recent returns and news, high scores on literacy / knowledge and few inconsistencies). In between, the "extrapolators" and "mean reverters" expect returns of around zero, have average literacy scores and errors, but they differ sharply in their reaction to returns and news. The extrapolators expect recent trends of both to continue, whereas mean reverters think that the opposite will happen. Finally, the group of "ignorants" stands out from the rest in that they do not seem to be very interested in financial matters, which results in frequent fifty-fifty answers to probabilistic expectations questions. On an ensuing question about whether these answers are supposed to express actual probabilistic judgements or general epistemic uncertainty, they often state the latter. Beliefs and their heterogeneous trajectories are reflected in predicted trading patterns. Our results are robust to different modelling assumptions in a number of directions. None of the five groups passes a rational expectations test; they all overreact in one way or another to recent information.

The evidence that households' expectations about the development of the stock market are heterogenous is overwhelming; Giglio, Maggiori, Stroebel, and Utkus (2019) is a recent contribution and contains a good overview of previous studies. We have shown that part of this can be traced to heterogenous expectations formations processes. In particular, the much longer time series has allowed us to go beyond the early contribution by Dominitz and Manski (2011) and classify individuals based on a statistical algorithm as opposed to inferring it from two observations only. An important step will be, of course, to replicate our findings in other datasets. Whereas the groups we could identify were fairly stable in our setting, it would be important to see whether comparable findings emerge from other question formats and in other countries. If so, this would have important implications for explaining stock market participation and for asset pricing models. For example, Barberis, Greenwood, Jin, and Shleifer (2015) develop an asset pricing model with extrapolative investors in addition to rational market participants. Our results suggest that even more investor types deserve such attention.

Appendix 1.A Data and stylised facts

1.A.1 Additional details

We use several background variables from the ALP, which are available for everyone. These include age, sex, education, and income. The other variables were regularly measured as part of the "Effects of the Financial Crisis" survey waves, or they come from other surveys in the ALP. A detailed source for each variable is given below so all details can be retrieved from https://alpdata.rand.org/index.php?page=data. The main differences are that we combine the two survey identifiers to be found for waves 16 (ALP survey identifiers 129 and 131) and 44 (288, 293) of the Effects of the Financial Crisis survey and that we display the number of observations for each variable that we can effectively use. The number of belief measures per wave is reduced substantially midway through the sample because a second and hard-to-compare format for belief measurement was introduced; which format was shown to individuals was drawn randomly anew in each wave.

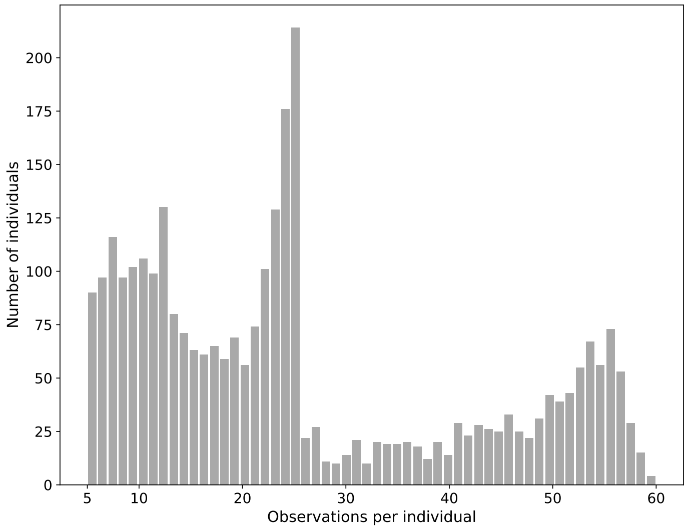


Figure 1.A.1. Distribution of the number of belief measurements per individual

Table 1.A.1. Details on surveys and variables

ID	Survey Title	Fielded	Variable Name	N
57	Effects of the Financial Crisis [W01]	2008-11	Beliefs	1840
			Follows/Understands stock market	1867
63	Effects of the Financial Crisis [W02]	2009-02	Beliefs	1710
			Follows/Understands stock market	1970
64	Financial Literacy March 09	2009-03	Financial Numeracy Financial Knowledge	1564
74	Effects of the Financial Crisis [W03]	2009-05	Beliefs	1796
			Owns stocks	1978
83	Effects of the Financial Crisis [W04]	2009-06	Beliefs	1876
			Owns stocks	2017
85	Effects of the Financial Crisis [W05]	2009-07	Beliefs	1904
			Owns stocks	2034
88	Effects of the Financial Crisis [W06]	2009-08	Beliefs	1867
			Owns stocks	2026
90	Effects of the Financial Crisis [W07]	2009-09	Beliefs	1901
			Owns stocks	2060
92	Effects of the Financial Crisis [W08]	2009-10	Beliefs	1816
			Owns stocks	1971
97	Effects of the Financial Crisis [W09]	2009-11	Beliefs	1891
			Owns stocks	2022
103	Effects of the Financial Crisis [W10]	2009-12	Beliefs	1904
			Owns stocks	2036
107	Effects of the Financial Crisis [W11]	2010-01	Beliefs	1879
			Owns stocks	2032
			Follows/Understands stock market	2040
111	Effects of the Financial Crisis [W12]	2010-02	Beliefs	1890

Continued on next page

ID	Survey Title	Fielded	Variable Name	N
116	Effects of the Financial Crisis [W13]	2010-03	Owns stocks	2023
			Beliefs	1862
117	Effects of the Financial Crisis [W14]	2010-04	Owns stocks	1996
			Beliefs	1798
124	Effects of the Financial Crisis [W15]	2010-05	Owns stocks	1927
			Follows/Understands stock market	1934
			Beliefs	1720
129	Effects of the Financial Crisis [W16]	2010-06	Owns stocks	1852
			Beliefs	1775
134	Effects of the Financial Crisis [W17]	2010-07	Owns stocks	1911
			Beliefs	1668
139	Effects of the Financial Crisis [W18]	2010-08	Owns stocks	1793
			Beliefs	1640
152	Effects of the Financial Crisis [W19]	2010-09	Owns stocks	1741
			Beliefs	1695
157	Effects of the Financial Crisis [W20]	2010-10	Owns stocks	1817
			Beliefs	1659
158	Effects of the Financial Crisis [W21]	2010-11	Owns stocks	1770
			Beliefs	1706
161	Effects of the Financial Crisis [W22]	2010-12	Owns stocks	1830
			Beliefs	1726
162	Effects of the Financial Crisis [W23]	2011-01	Owns stocks	1848
			Beliefs	1666
173	Effects of the Financial Crisis [W24]	2011-02	Owns stocks	1801
			Beliefs	1722
176	Effects of the Financial Crisis [W25]	2011-03	Owns stocks	1812
			Follows/Understands stock market	1815
			Beliefs	1708
			Owns stocks	1828

Continued on next page

ID	Survey Title	Fielded	Variable Name	N
178	Effects of the Financial Crisis [W26]	2011-04	Beliefs	860
			Owns stocks	1759
188	Effects of the Financial Crisis [W27]	2011-05	Beliefs	859
			Owns stocks	1748
194	Effects of the Financial Crisis [W28]	2011-06	Beliefs	872
			Owns stocks	1749
198	Effects of the Financial Crisis [W29]	2011-07	Beliefs	889
			Owns stocks	1805
208	Effects of the Financial Crisis [W30]	2011-08	Beliefs	883
			Owns stocks	1820
211	Effects of the Financial Crisis [W31]	2011-09	Beliefs	846
			Owns stocks	1759
219	Effects of the Financial Crisis [W32]	2011-10	Beliefs	826
			Owns stocks	1746
225	Effects of the Financial Crisis [W33]	2011-11	Beliefs	937
			Owns stocks	1792
231	Effects of the Financial Crisis [W34]	2011-12	Beliefs	898
			Owns stocks	1742
236	Effects of the Financial Crisis [W35]	2012-01	Beliefs	960
			Owns stocks	1796
239	Effects of the Financial Crisis [W36]	2012-02	Beliefs	955
			Owns stocks	1795
249	Effects of the Financial Crisis [W37]	2012-03	Beliefs	906
			Owns stocks	1682
253	Effects of the Financial Crisis [W38]	2012-04	Beliefs	946
			Owns stocks	1756
			Follows/Understands stock market	1757
262	Effects of the Financial Crisis [W39]	2012-05	Beliefs	805
			Owns stocks	1507

Continued on next page

ID	Survey Title	Fielded	Variable Name	N
267	Effects of the Financial Crisis [W40]	2012-06	Beliefs	892
			Owns stocks	1677
271	Effects of the Financial Crisis [W41]	2012-07	Beliefs	928
			Owns stocks	1743
			Follows/Understands stock market	1754
278	Effects of the Financial Crisis [W42]	2012-08	Beliefs	927
			Owns stocks	1698
281	Effects of the Financial Crisis [W43]	2012-09	Beliefs	907
			Owns stocks	1669
288	Effects of the Financial Crisis [W44]	2012-10	Beliefs	1023
			Owns stocks	1820
299	Effects of the Financial Crisis [W45]	2012-11	Beliefs	1312
			Owns stocks	2181
305	Effects of the Financial Crisis [W46]	2012-12	Beliefs	1347
			Owns stocks	2204
			Knows stock return from last year	1039
322	Effects of the Financial Crisis [W47]	2013-01	Beliefs	1101
			Owns stocks	1720
328	Effects of the Financial Crisis [W48]	2013-02	Beliefs	1317
			Owns stocks	2086
			Knows stock return from last year	1008
332	Effects of the Financial Crisis [W49]	2013-03	Beliefs	1300
			Owns stocks	2145
335	Effects of the Financial Crisis [W50]	2013-04	Beliefs	1347
			Owns stocks	2100
			Follows/Understands stock market	1467
			Knows stock return from last year	1176

Continued on next page

ID	Survey Title	Fielded	Variable Name	N
345	Effects of the Financial Crisis [W51]	2013-07	Beliefs	98
			Owns stocks	157
358	Effects of the Financial Crisis [W52]	2013-10	Beliefs	881
			Owns stocks	1428
363	Reasons for expectations [W01]	2013-12	Reasons for expectations	114
368	Effects of the Financial Crisis [W53]	2014-01	Beliefs	939
			Owns stocks	1478
379	Effects of the Financial Crisis [W54]	2014-04	Beliefs	257
			Owns stocks	439
			Follows/Understands stock market	279
389	Effects of the Financial Crisis [W55]	2014-07	Beliefs	924
			Owns stocks	1500
			Follows/Understands stock market	1498
400	Effects of the Financial Crisis [W56]	2014-10	Beliefs	311
			Owns stocks	540
417	Effects of the Financial Crisis [W57]	2015-01	Beliefs	949
			Owns stocks	1547
426	Effects of the Financial Crisis [W58]	2015-04	Beliefs	794
			Owns stocks	1303
			Follows/Understands stock market	1296
			Probability Numeracy	1291
434	Effects of the Financial Crisis [W59]	2015-07	Beliefs	825
			Owns stocks	1323
			Follows/Understands stock market	1320
440	Effects of the Financial Crisis [W60]	2015-10	Beliefs	813
			Owns stocks	1340
			Follows/Understands stock market	1335
			Probability Numeracy	1310

Continued on next page

ID	Survey Title	Fielded	Variable Name	N
448	Effects of the Financial Crisis [W61]	2016-01	Beliefs	1057
			Owns stocks	1700
			Follows/Understands stock market	1693
			Probability Numeracy	1670

1.A.2 Detailed stylised facts (Section 1.2.2)

1.A.2.1 On average, beliefs are pessimistic compared to historical returns

A comparison of the average subjective beliefs with the distribution of historical returns reveals that the individuals in our sample are pessimistic about the stock market. This finding is in line with Hurd's 2009 summary of various studies and data as well as Hurd, Rooij, and Winter's 2011 report for Dutch households.

In Table 1.A.2 we collected expected returns and probabilities for returns exceeding -20%, 0% and 20% from the historical data and compare them with the average subjective beliefs. Individuals are too pessimistic by 23 and 28 percentage points respectively that the Dow Jones will not collapse and that it will increase. The fact that individuals seem to be too optimistic that the Dow Jones will increase by 20 percent or more relative to empirical frequencies should probably not be taken at face value. If we drop individuals who exhibit monotonicity violations from the sample, the difference changes sign in line with the other values. In sum, relative to the historical distribution, individuals are, on average, too pessimistic.

Table 1.A.2. Historical returns vs. beliefs about returns

	Historical Averages	Subjective Beliefs	Difference
$E[R_{t\to t+12}]$	7.3	0.5	6.9
$\Pr(R_{t\to t+12}) > -0.2$	97.1	74.6	22.5
$\Pr(R_{t\to t+12}) > 0$	72.1	44.0	28.1
$\Pr(R_{t\to t+12}) > 0.2$	23.5	26.8	-3.3

Notes: Units in percentage points. The historical averages $\Pr(R_{t\to t+12} > x)$ are estimated using the empirical frequency $T^{-1}\sum_t^T 1\{R_{t\to t+12} > x\}$ for yearly returns of the Dow Jones between 1950 and 2016. Beliefs are within-person means.

1.A.2.2 Beliefs exhibit significant dispersion within and across individuals

Table 1.2 has already shown the substantial variation in average beliefs across individuals. The same holds true for the variation within persons across time with comparable magnitudes (see Table 1.A.3).

Table 1.A.3. Within- and between-person variation of belief variables

	Average within-subject std.	Between-subject std.	Ratio
$E[R_{t \to t+12}]$	5.5	5.8	0.94
$\Pr(R_{t \to t+12}) > -0.2$	13.3	13.4	1.00
$\Pr(R_{t \to t+12}) > 0$	15.5	17.8	0.87
$\Pr(R_{t \to t+12}) > 0.2$	14.6	14.2	1.03

Notes: Units in percentage points.

One notable feature is that for both within and between-subject differences, the variation is largest for the first and arguably most intuitive question, i.e., $\Pr(R_{t \to t+12} > 0)$. In the next subsection, we confirm that individual characteristics have most predictive power for variation in this measure of an individual's beliefs about the future of the Dow Jones.

Figure 1.A.2 shows that the substantial belief variation over time we find at the individual level largely cancels out if beliefs are averaged across subjects. Unless the within-variation is unsystematic, this is an indication that average beliefs averages mask substantial heterogeneity in belief dynamics.

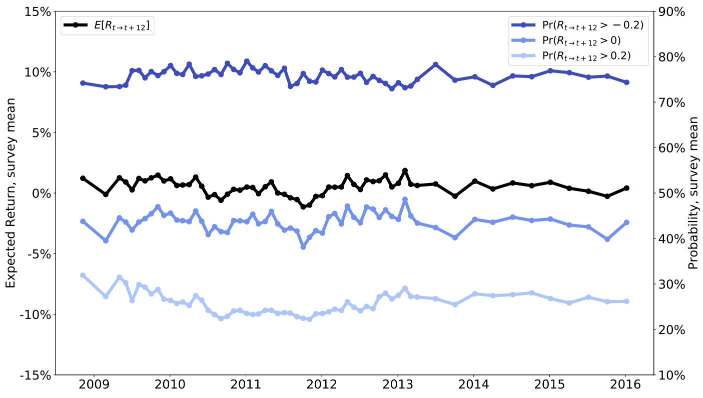


Figure 1.A.2. Average beliefs over time

Notes: Depicted series are within-survey means. The left y-axis displays the scale for the expected returns, the right y-axis displays the scale for the three probabilities.

1.A.2.3 Beliefs of financially sophisticated and knowledgeable individuals are more optimistic

To get a sense of what is driving persistent level differences in beliefs, we once more average beliefs within individuals and then regress them on individual-level characteristics. Table 1.A.2.3 reports the results.

Table 1.A.4. Predictors of average beliefs, all regressors

	$E[R]$	$\Pr(R > -0.2)$	$\Pr(R > 0)$	$\Pr(R > 0.2)$
Follows stock market	1.16**	0.29	3.56**	1.61
	(0.53)	(1.23)	(1.61)	(1.30)
Understands stock market	0.41	0.02	2.12	0.00
	(0.51)	(1.13)	(1.50)	(1.22)
Knowl. of past returns: Don't know	1.45**	1.31	4.30**	1.41
	(0.61)	(1.56)	(1.88)	(1.61)
Knowl. of past returns: Magnitude too large	4.84***	2.01	13.35***	8.12***
	(1.07)	(2.15)	(2.70)	(2.73)
Knowl. of past returns: Sign and Mag. correct	3.06***	3.26**	9.23***	2.71*
	(0.62)	(1.47)	(1.87)	(1.54)
Probability Numeracy	1.09***	0.22	4.07***	0.99
	(0.26)	(0.56)	(0.78)	(0.67)
Financial Knowledge	0.34	0.71	2.09***	-1.12
	(0.25)	(0.64)	(0.81)	(0.71)
Financial Numeracy	0.12	0.37	1.19	-0.94
	(0.25)	(0.57)	(0.76)	(0.67)
Intercept	-10.25***	95.85***	-6.74	11.85
	(3.80)	(9.76)	(12.72)	(10.87)
Age	-0.02	-0.94***	0.20	0.36
	(0.09)	(0.23)	(0.28)	(0.25)
Age squared	0.00	0.01***	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Male	0.01	-1.26	2.21*	-1.45
	(0.43)	(0.95)	(1.25)	(1.08)
Education: Some college	0.69	-2.72*	2.15	2.88*
	(0.60)	(1.56)	(1.87)	(1.56)
Education: Bachelor degree	1.69***	-0.96	5.75***	3.10*
	(0.65)	(1.67)	(1.97)	(1.66)
Education: Advanced degree	2.71***	0.71	8.68***	3.47*
	(0.73)	(1.63)	(2.16)	(1.81)
Ethnicity: Black	1.32	1.43	4.21	1.50
	(1.80)	(3.62)	(4.92)	(5.32)
Ethnicity: Native	-0.36	-28.22***	10.16**	4.08
	(1.76)	(3.68)	(4.86)	(5.14)

Continued on next page

Table 1.A.4. Predictors of average beliefs, all regressors

	$E[R]$	$\Pr(R > -0.2)$	$\Pr(R > 0)$	$\Pr(R > 0.2)$
Ethnicity: Other	-1.30	-4.70	0.02	-2.94
	(1.87)	(6.59)	(5.71)	(5.43)
Ethnicity: White	0.47	0.02	4.68	-2.28
	(1.52)	(3.03)	(3.98)	(4.68)
Household income (thousands), $\in (5, 7.5]$	3.04	-6.33	9.79	9.55
	(4.55)	(9.24)	(16.47)	(10.74)
Household income (thousands), $\in (7.5, 10]$	-0.28	0.50	0.99	-2.69
	(3.08)	(7.96)	(10.74)	(7.98)
Household income (thousands), $\in (10, 12.5]$	-1.86	-12.31	-1.27	2.30
	(3.36)	(8.31)	(10.93)	(8.04)
Household income (thousands), $\in (12.5, 15]$	0.25	-0.88	1.65	0.39
	(3.03)	(7.38)	(11.15)	(7.82)
Household income (thousands), $\in (15, 20]$	1.24	-5.98	4.04	6.73
	(3.00)	(7.77)	(10.86)	(7.86)
Household income (thousands), $\in (20, 25]$	0.85	-6.14	3.15	5.15
	(2.89)	(7.15)	(10.54)	(7.67)
Household income (thousands), $\in (25, 30]$	0.23	-0.23	1.28	0.61
	(2.89)	(7.26)	(10.76)	(7.58)
Household income (thousands), $\in (30, 35]$	0.32	-4.00	1.07	3.99
	(2.86)	(7.04)	(10.51)	(7.47)
Household income (thousands), $\in (35, 40]$	2.50	-0.97	7.33	5.29
	(2.84)	(7.03)	(10.50)	(7.48)
Household income (thousands), $\in (40, 50]$	0.64	-3.00	3.41	2.12
	(2.79)	(6.88)	(10.31)	(7.32)
Household income (thousands), $\in (50, 60]$	1.13	-2.39	4.74	2.90
	(2.82)	(6.93)	(10.37)	(7.42)
Household income (thousands), $\in (60, 75]$	0.93	-3.16	3.00	4.15
	(2.81)	(6.85)	(10.33)	(7.34)
Household income (thousands), $\in (75, 100]$	1.73	-0.83	5.49	3.40
	(2.78)	(6.80)	(10.29)	(7.25)
Household income (thousands), $\in (100, 125]$	0.64	-2.41	3.98	0.97
	(2.79)	(6.81)	(10.34)	(7.28)
Household income (thousands), $\in (125, 200]$	0.76	-0.57	2.21	2.24
	(2.84)	(6.82)	(10.39)	(7.40)
Household income (thousands), $> 200k$	-0.92	-2.17	-1.78	-0.45
	(3.01)	(7.15)	(10.75)	(7.68)
Ever owned stocks	-0.29	-1.16	-1.03	0.86
	(0.57)	(1.21)	(1.57)	(1.37)
N	805	805	805	805
R^2	0.22	0.11	0.32	0.05

OLS estimates. Standard errors (robust) in parentheses. ***, ** and * denote significance at 1%, 5% and 10% respectively. Omitted categories are 'Does not follow stock market', 'Does not understand stock market', and 'Knowledge of past return: Wrong sign given'. Dependent variables are within-person means in percentage points. Measures of Financial and probability numeracy are standardised.

We focus on variables that capture the extent to which people are involved with, and have knowledge of, the stock market and financial matters more generally. All regressions included controls. The signs of the significant predictors confirm what we would expect: A better knowledge of past returns and financial matters, as well as following the stock market, are associated with more optimistic beliefs. Meanwhile, self-assessed understanding does help much to predict beliefs conditional on knowledge of past returns and financial knowledge.

Knowledge of past returns, our most direct measure of an individuals' information set, is the strongest predictor for expected returns and for all three probabilistic beliefs. Relative to respondents who state the wrong sign for a Dow Jones return over the past year, individuals who give the correct sign and magnitude (or overestimate the latter) are 9-13 percentage points more optimistic that the Dow will increase over the coming year; they expect returns that are 3-5 percentage points higher on average.

Higher probability numeracy and financial knowledge also predict optimism in the belief that the Dow will increase. A one standard deviation increase in these scores predicts increases in the beliefs that the Dow will rise of 4 and 2 percentage points. That probability numeracy is associated with belief levels conditional on various indicators measuring what people know about the stock market points at measurement error in stated beliefs.

As noted before, the predictive power of the covariates is much higher for the probability of a positive return than for the other two points on the distribution function; the R^2 differs by factors of three to five. We take this as additional evidence pointing towards higher noise levels for the events of the Dow Jones rising or falling by at least 20%. Put differently, we should not take all stated beliefs at face value.

1.A.2.4 Stated beliefs vary in their information value

Measurement error and / or imprecision in stated beliefs have concerned researchers for a long time. Two particularly prevalent phenomena are rounding of stated probabilities and the previously-mentioned monotonicity violations. We regard both as indications that stated measures are less informative about what an individual thinks about the stock market, similar in spirit to Drerup, Enke, and von Gaudecker (2017).

Figure 1.A.3 shows histograms of the beliefs with 1-percent bins. Most beliefs are rounded to the nearest multiple of 5% or 10%, and that answers equalling 50% are particularly frequent. The middle Panel of Figure 1.A.3 looks very similar to Figure 3 in Hurd, Rooij, and Winter (2011). These basic facts on rounding have been documented for a long time, Manski and Molinari (2010) and Kleinjans and Soest (2014) are recent contributions and modelling suggestions. Rounding suggests indi-

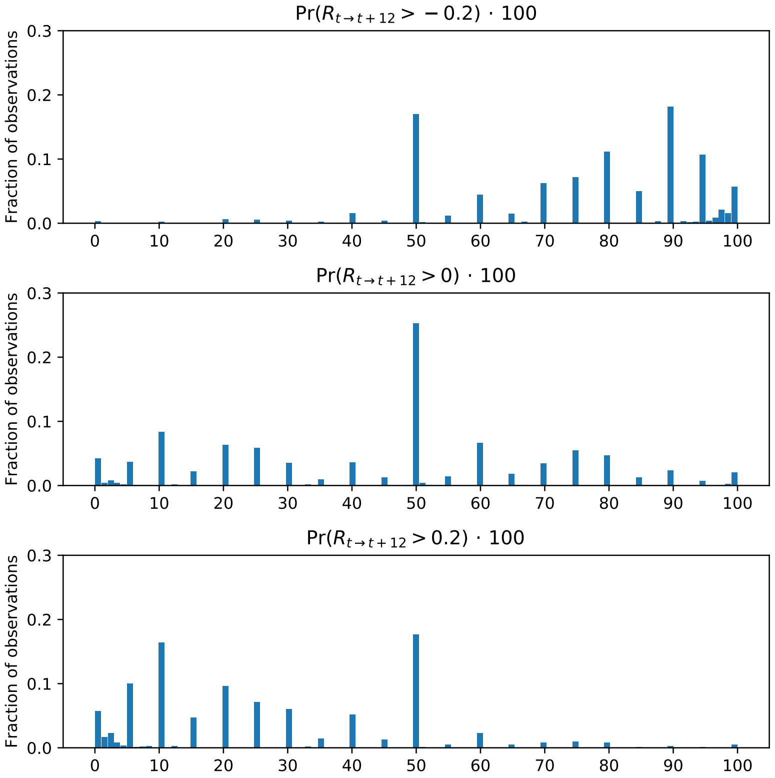


Figure 1.A.3. Distributions of belief variables

Notes: The figures depict histograms of belief variables with 1-percent bins. Data is pooled across surveys, including only individuals with at least five sets of belief measurements.

viduals are either not willing to exert the effort to express a precise belief, or that their beliefs themselves are imprecise.

Bruin, Fischhoff, Millstein, and Halpern-Felsher (2000) argue that 50% answers might indicate that individuals are epistemically uncertain about an event rather than expressing subjective beliefs of equal likelihoods. Following up on that observation, the questionnaires that we use confront respondents who gave an answer equal to 50% for $\Pr(R_{t\to t+12} \leq 0)$ with a follow up question. It asks them to clarify whether they mean that the Dow Jones is equally likely to rise as it is to fall, or whether they are simply unsure. 47% of responses to this question indicated that they are unsure, not that they judge the probabilities to be equal. As one would expect if people do not have a well formed belief, the stated probability for $\Pr(R_{t\to t+12} \geq 0.2)$ and $\Pr(R_{t\to t+12} \leq -0.2)$ also equalled 50% about half of the time in in that case. By contrast, for the 53% of responses indicating that a probability of 50% means they find an increase and a decrease equally likely, the other two probabilities equalled 50% only one third of the time.

A striking irregularity in measured beliefs are monotonicity violations. Similarly to rounding, this is in line with what previous studies of probabilistic expectations have found (e.g. Hudomiet, Kezdi, and Willis, 2010; Hurd, Rooij, and Winter, 2011). Our raw beliefs data consists of 3 points on the cumulative distribution function: −0.2, 0 and 0.2. There was no reminder that stated beliefs have to (weakly) increase along these points, and hence answers can violate the monotonicity property of the cumulative distribution function. Stated beliefs that are not monotone are incoherent, and thus cannot be regarded as very informative about what people believe will happen with the Dow Jones. To a somewhat lesser extent, this is true for weakly but not strongly monotone beliefs as well. While compatible with probability calculus, such answers suggest respondents think there is no chance the return of the Dow could be between -20% and 0% or 0% and 20%, even as they do think there is a chance returns could be smaller or larger than that. Table 1.A.5 shows the incidence of monotonicity violations in our data. Around 70% of stated beliefs sets are strictly monotone between the points −0.2 and 0 as well as 0 and 0.2, making for 57% that satisfy both checks.

Table 1.A.5. Prevalence of monotonicity violations

	From -0.2 to 0	From 0 to 0.2	Either
Not monotone	0.08	0.07	0.15
Weakly but not strictly monotone	0.18	0.23	0.28
Strictly monotone	0.74	0.70	0.57

Notes: Table shows fraction of beliefs satisfying each listed monotonicity status.

Table 1.A.5 shows that a substantial number of people give answers that do not obey the rules of probability calculus or seem implausible. The propensity to give monotonicity violating answers may be thought of as being determined by the effort give when answering the survey and by how much effort is required to avoid errors and give reasonable answers. While we cannot observe effort, people familiar with financial markets, in particular stock markets, should find it easier to avoid mistakes. In addition, such people are more likely to hold precise beliefs in the first place, as their information set is richer. Knowledge of probability calculus and familiarity with using probabilities to indicate uncertainty can also be expected to reduce the incidence of nonsensical answers. Both of these are likely positively related to effort, as people are more willing to do tasks they are good at and interested in.

1.A.2.5 Beliefs of financially sophisticated and knowledgeable individuals are more consistent

To investigate what drives monotonicity violations, epistemic uncertainty, and rounding we use measures of probability numeracy, financial numeracy and engagement with the stock market along with typical characteristics such as gender, age, education, income and ethnicity. As before, we collapse the time dimension of our data. We compute an individual's average propensity to express non-monotone or weakly monotone beliefs, their average propensity to say that their 50% beliefs mean they are unsure as opposed to a subjective probability (if individuals did not see this follow up question because they did not give a 50% answer, we assume their answer is a subjective probability) and their average propensity to give answers that are multiples of 5% as dependent variables. We regress these on personal characteristics. Kezdi and Willis (2008) and Gouret and Hollard (2011) find no relationship between the propensity to give problematic answers and general personal characteristics, but we find strong relationships between financial and probability numeracy and non-monotone or epistemically uncertain beliefs.

Table 1.A.6. Predictors of non-monotonicity, epistimic uncertainty, and rounding

	Non-monotone	Epistemically unsure	Rounded to 10%
Follows stock market	-0.05**	-0.02	0.02
	(0.02)	(0.01)	(0.02)
Understands stock market	-0.03	-0.02**	-0.02
	(0.02)	(0.01)	(0.02)
Knowl. of past returns: Don't know	-0.07**	0.01	-0.01
	(0.03)	(0.02)	(0.03)
Knowl. of past returns: Magnitude too large	-0.05	0.01	-0.02
	(0.04)	(0.03)	(0.04)
Knowl. of past returns: Sign and Mag. correct	-0.08***	-0.01	-0.04
	(0.03)	(0.02)	(0.03)
Probability Numeracy	-0.06***	-0.01	0.01
	(0.01)	(0.01)	(0.01)
Financial Knowledge	-0.06***	-0.03***	-0.01
	(0.01)	(0.01)	(0.01)
Financial Numeracy	-0.04***	0.00	0.00
	(0.01)	(0.01)	(0.01)
Intercept	0.59***	0.18	0.66***
	(0.17)	(0.11)	(0.22)
Age	0.01***	-0.00	-0.00
	(0.00)	(0.00)	(0.00)
Age squared	-0.00***	0.00	0.00
	(0.00)	(0.00)	(0.00)
Male	-0.02	0.00	-0.01
	(0.02)	(0.01)	(0.02)
Education: Some college	0.01	0.01	0.02
	(0.03)	(0.02)	(0.03)
Education: Bachelor degree	-0.05*	0.01	0.04
	(0.03)	(0.02)	(0.03)

Continued on next page

Table 1.A.6. Predictors of non-monotonicity, epistimic uncertainty, and rounding

	Non-monotone	Epistemically unsure	Rounded to 10%
Education: Advanced degree	-0.06*	-0.01	-0.02
	(0.03)	(0.02)	(0.03)
Ethnicity: Black	0.04	-0.02	0.10
	(0.08)	(0.04)	(0.07)
Ethnicity: Native	0.31***	-0.11***	-0.01
	(0.08)	(0.04)	(0.07)
Ethnicity: Other	0.00	0.00	0.17
	(0.11)	(0.07)	(0.11)
Ethnicity: White	-0.10	-0.01	0.08
	(0.07)	(0.03)	(0.05)
Household income (thousands), $\in (5, 7.5]$	-0.09	-0.01	0.18
	(0.15)	(0.08)	(0.20)
Household income (thousands), $\in (7.5, 10]$	-0.07	0.01	0.07
	(0.11)	(0.10)	(0.17)
Household income (thousands), $\in (10, 12.5]$	0.09	0.05	0.11
	(0.11)	(0.09)	(0.18)
Household income (thousands), $\in (12.5, 15]$	-0.05	-0.02	-0.02
	(0.11)	(0.08)	(0.17)
Household income (thousands), $\in (15, 20]$	0.00	-0.03	0.08
	(0.11)	(0.08)	(0.17)
Household income (thousands), $\in (20, 25]$	-0.01	0.04	0.13
	(0.10)	(0.08)	(0.16)
Household income (thousands), $\in (25, 30]$	-0.05	-0.05	-0.00
	(0.10)	(0.07)	(0.16)
Household income (thousands), $\in (30, 35]$	0.01	0.02	0.09
	(0.10)	(0.08)	(0.16)
Household income (thousands), $\in (35, 40]$	-0.09	-0.04	0.03
	(0.10)	(0.07)	(0.16)
Household income (thousands), $\in (40, 50]$	-0.03	-0.00	0.08
	(0.09)	(0.07)	(0.16)
Household income (thousands), $\in (50, 60]$	-0.07	0.02	0.09
	(0.09)	(0.08)	(0.16)
Household income (thousands), $\in (60, 75]$	-0.02	-0.01	0.06
	(0.09)	(0.07)	(0.16)
Household income (thousands), $\in (75, 100]$	-0.09	-0.01	0.06
	(0.09)	(0.07)	(0.16)
Household income (thousands), $\in (100, 125]$	-0.08	-0.03	0.02
	(0.09)	(0.07)	(0.16)
Household income (thousands), $\in (125, 200]$	-0.10	-0.03	0.10
	(0.09)	(0.07)	(0.16)
Household income (thousands), > 200	-0.08	-0.02	0.13

Continued on next page

Table 1.A.6. Predictors of non-monotonicity, epistimic uncertainty, and rounding

	Non-monotone	Epistemically unsure	Rounded to 10%
	(0.10)	(0.08)	(0.16)
Ever owned stocks	0.07***	0.04***	0.04*
	(0.02)	(0.01)	(0.02)
$\Pr(R_{t\rightarrow t+12} > 0)$	0.49***	0.11***	-0.34***
	(0.05)	(0.03)	(0.06)
N	805	805	805
R^2	0.31	0.13	0.13

OLS estimates. Standard errors (robust) in parentheses. ***, ** and * denote significance at 1%, 5% and 10% respectively. Omitted categories are 'Does not follow stock market', 'Does not understand stock market', and 'Knowledge of past return: Wrong sign given'. Measures of financial and probability numeracy are standardised.

In line with the earlier discussion, the regression results in Table 1.A.2.5 demonstrate that following the stock market, having accurate knowledge of historical returns, probability numeracy, and financial numeracy all predict that an individual is less likely to express beliefs afflicted by monotonicity errors. The most important predictors for individuals to state that their expressed beliefs indicate likelihoods are self-assessed understanding of the stock market, probability numeracy and financial numeracy. One interpretation of these associations is that richer information sets and greater understanding lead to more precise beliefs, and lower the costs of stating beliefs in the survey, which reduces the incidence of errors. Greater familiarity with probabilities also lowers errors and makes it more likely that individuals use 50% answers to indicate equal likelihoods. Rounding, measured as the fraction of answers that are multiples of 5, is not systematically predictable with our indicators of sophistication.

Appendix 1.B Additional details for the main model

Table 1.B.1. Group sizes

	N	Share of sample
Pessimists	749	0.25
Mean Reverters	584	0.19
Extrapolators	527	0.17
Ignorants	393	0.13
Sophisticates	777	0.26
Total	3030	

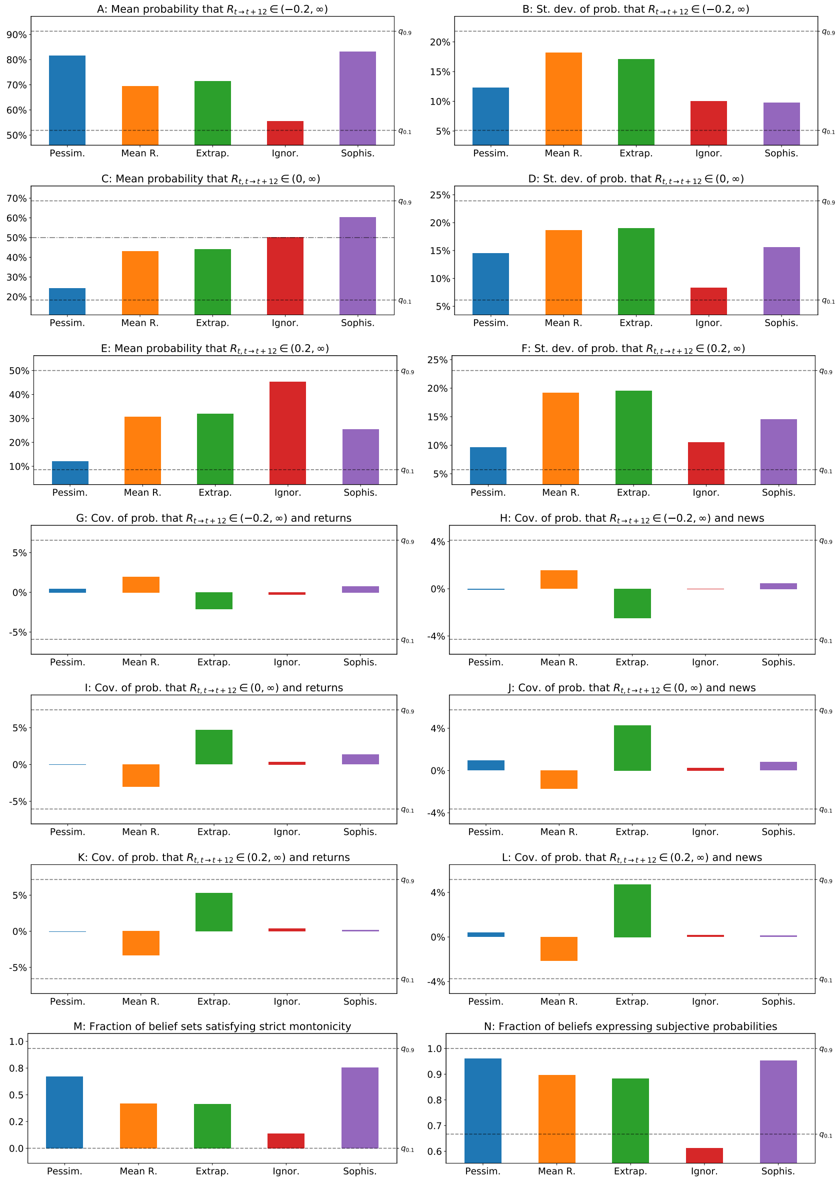


Figure 1.B.1. Moments used for classification by unobserved heterogeneity group

Notes: Bars show the group means of the 14 individual moments used to classify individuals via the k-means algorithm. Dashed lines are the bottom and top decile with respect to the individuals of all groups taken together.

Table 1.B.2. Coefficients for main specification when L = 0

	Pessim.	Mean R.	Extrap.	Ignor.	Sophis.
Intercept	-4.74***	-0.45	3.06***	2.93***	5.32***
	(0.15)	(0.32)	(0.32)	(0.21)	(0.18)
Lag 0, Returns	-0.02	-0.53***	0.72***	0.12**	0.24***
	(0.03)	(0.06)	(0.07)	(0.05)	(0.04)
Lag 0, News	0.42***	-0.47***	1.15***	0.07	0.27***
	(0.05)	(0.10)	(0.10)	(0.06)	(0.05)
N · T			77310		
R^2			0.256		

Notes: N = 3030. Individuals for whom not all covariates are available are excluded. OLS estimates. Standard errors (clustered at individual level) in parentheses. ***, ** and * denote significance at 1%, 5% and 10% respectively. Dependent variable in percentage points, regressors standardised.

Table 1.B.3. Coefficients for main specification when L = 6

	Pessim.	Mean R.	Extrap.	Ignor.	Sophis.
Intercept	-3.35***	-0.42	3.73***	3.26***	6.22***
	(0.28)	(0.61)	(0.60)	(0.35)	(0.33)
Lag 0, Returns	0.04	-0.55***	0.62***	0.09*	0.22***
	(0.04)	(0.07)	(0.07)	(0.05)	(0.04)
Lag 1, Returns	-0.02	0.04	0.41***	0.17***	0.15***
	(0.04)	(0.07)	(0.07)	(0.05)	(0.04)
Lag 2, Returns	0.01	-0.09	-0.00	0.16***	0.10***
	(0.03)	(0.06)	(0.07)	(0.05)	(0.04)
Lag 3, Returns	-0.01	-0.05	0.11*	0.05	0.12***
	(0.03)	(0.06)	(0.07)	(0.04)	(0.03)
Lag 4, Returns	-0.08***	0.01	0.10*	0.06*	0.08**
	(0.03)	(0.05)	(0.06)	(0.04)	(0.03)
Lag 5, Returns	-0.08***	-0.10*	-0.01	0.08**	0.01
	(0.03)	(0.06)	(0.06)	(0.03)	(0.03)
Lag 6, Returns	-0.08***	0.05	0.00	-0.01	0.02
	(0.03)	(0.06)	(0.07)	(0.04)	(0.03)
Lag 0, News	0.32***	-0.45***	1.04***	0.12*	0.17***
	(0.05)	(0.10)	(0.09)	(0.06)	(0.05)
Lag 1, News	0.13***	-0.03	0.13	-0.26***	-0.02
	(0.05)	(0.11)	(0.10)	(0.07)	(0.05)
Lag 2, News	0.23***	0.33***	0.14	0.07	0.31***
	(0.05)	(0.12)	(0.12)	(0.06)	(0.06)
Lag 3, News	0.16***	-0.04	0.41***	0.07	0.06
	(0.05)	(0.12)	(0.13)	(0.08)	(0.06)
Lag 4, News	-0.03	-0.26**	-0.10	0.08	0.05
	(0.06)	(0.12)	(0.13)	(0.08)	(0.07)
Lag 5, News	0.07	0.10	-0.24*	0.06	0.04
	(0.06)	(0.13)	(0.13)	(0.08)	(0.07)
Lag 6, News	0.19***	-0.09	0.14	0.14*	0.13**
	(0.06)	(0.13)	(0.13)	(0.08)	(0.06)
N · T			77310		
R^2			0.256		

Notes: N = 3030. OLS estimates. Standard errors (clustered at individual level) in parentheses. ***, ** and * denote significance at 1%, 5% and 10% respectively. Dependent variable in percentage points, regressors standardised.

Table 1.B.4. Within-group heterogeneity

		Main specifciation	Fin. Know.: > med	Fin. Num.: > med	Underst. stock m.	Follows stock m.	Know. past ret.	Age: > med
Pessim.	Returns	-0.04	0.31***	0.28***	0.30***	0.24**	0.25***	0.02
		(0.05)	(0.10)	(0.10)	(0.10)	(0.09)	(0.10)	(0.11)
	News	0.47***	-0.02	-0.25*	-0.16	-0.16	-0.16	0.10
		(0.07)	(0.15)	(0.15)	(0.14)	(0.14)	(0.14)	(0.16)
Mean R.	Returns	-0.52***	0.06	-0.05	0.05	-0.18	0.16	-0.05
		(0.09)	(0.20)	(0.19)	(0.19)	(0.18)	(0.19)	(0.20)
	News	-0.14	0.00	-0.21	-0.11	0.20	-0.31	0.26
		(0.15)	(0.33)	(0.30)	(0.32)	(0.27)	(0.30)	(0.37)
Extrap.	Returns	0.55***	0.45**	0.30	0.37*	0.63***	0.11	-0.01
		(0.11)	(0.20)	(0.22)	(0.22)	(0.20)	(0.21)	(0.24)
	News	1.14***	-0.30	0.20	-0.27	-0.18	-0.11	0.36
		(0.15)	(0.31)	(0.31)	(0.31)	(0.30)	(0.30)	(0.41)
Ignor.	Returns	0.11	-0.28	-0.01	-0.19	-0.34**	-0.04	-0.20
		(0.08)	(0.18)	(0.23)	(0.17)	(0.16)	(0.17)	(0.14)
	News	0.03	-0.03	-0.21	-0.08	0.01	-0.11	0.10
		(0.09)	(0.21)	(0.25)	(0.20)	(0.18)	(0.18)	(0.26)
Sophis.	Returns	0.22***	-0.03	-0.16	-0.02	0.05	0.08	-0.07
		(0.05)	(0.11)	(0.10)	(0.11)	(0.12)	(0.11)	(0.11)
	News	0.33***	-0.37***	-0.12	-0.29**	-0.28*	0.01	0.21
		(0.06)	(0.13)	(0.12)	(0.13)	(0.16)	(0.13)	(0.15)
	N · T	37828	37828	37828	37828	37828	37828	37828
	R^2	0.26	0.26	0.27	0.26	0.27	0.26	0.26

Notes: Individuals for whom not all covariates are available are excluded. The first column reproduces the coefficients from our main specification for the subsample of individuals for which all covariates in the adjacent columns are available. The adjacent columns show the difference for each row's coefficient between individuals with and without the status given in the column header.

Table 1.B.5. Robustness to rounding

		Main specfciation	With indicators for rounding
Pessim.	Intercept	-4.74***	-3.94***
		(0.15)	(0.21)
	Returns	-0.02	-0.02
		(0.03)	(0.04)
	News	0.42***	0.42***
		(0.05)	(0.05)
Mean R.	Intercept	-0.45	0.27
		(0.32)	(0.37)
	Returns	-0.53***	-0.52***
		(0.06)	(0.06)
	News	-0.47***	-0.46***
		(0.10)	(0.09)
Extrap.	Intercept	3.06***	3.72***
		(0.32)	(0.37)
	Returns	0.72***	0.67***
		(0.07)	(0.07)
	News	1.15***	1.14***
		(0.10)	(0.10)
Ignor.	Intercept	2.93***	3.90***
		(0.21)	(0.31)
	Returns	0.12**	0.10**
		(0.05)	(0.05)
	News	0.07	0.09
		(0.06)	(0.06)
Sophis.	Intercept	5.32***	5.63***
		(0.18)	(0.21)
	Returns	0.24***	0.20***
		(0.04)	(0.04)
	News	0.27***	0.24***
		(0.05)	(0.05)
	Pr($R \leq -20\%$) divisible by 10%		0.54***
			(0.11)
	Pr($R \leq 0\%$) divisible by 10%		-3.33***
			(0.16)
	Pr($R \leq 20\%$) divisible by 10%		1.62***
			(0.11)
	N · T	77310	77310
	R^2	0.25	0.29

Notes: The divisibility variables are dummies equal to 1 if the subjective probability it refers to is divisible by 10%. The outcome of the regressions, the expected return, is based on all three subjective probabilities.

Appendix 1.C Alternative specifications

1.C.1 Only observed heterogeneity

Table 1.C.1. Model with no heterogeneity and observed heterogeneity

	$E[R]$, no heterogeneity	$E[R]$, obs. heterogeneity
Intercept	1.83***	-6.80*
	(0.21)	(3.70)
Lag 0, Returns	0.06**	0.03
	(0.03)	(0.04)
Lag 1, Returns	0.12***	0.08**
	(0.03)	(0.04)
Lag 2, Returns	0.02	0.05*
	(0.02)	(0.03)
Lag 3, Returns	0.04*	0.08**
	(0.02)	(0.03)
Lag 4, Returns	0.02	0.04
	(0.02)	(0.03)
Lag 5, Returns	-0.02	-0.01
	(0.02)	(0.03)
Lag 6, Returns	0.00	0.02
	(0.02)	(0.03)
Lag 0, News	0.25***	0.31***
	(0.03)	(0.05)
Lag 1, News	-0.03	-0.03
	(0.03)	(0.05)
Lag 2, News	0.22***	0.33***
	(0.04)	(0.05)
Lag 3, News	0.13***	0.08
	(0.04)	(0.06)
Lag 4, News	-0.05	-0.01
	(0.04)	(0.06)
Lag 5, News	0.00	-0.00
	(0.04)	(0.06)
Lag 6, News	0.12***	0.14**
	(0.04)	(0.06)
Age		-0.01
		(0.09)
Age squared		0.00
		(0.00)
Male		0.13
		(0.46)

Continued on next page

Table 1.C.1. Model with no heterogeneity and observed heterogeneity

	$E[R]$, no heterogeneity	$E[R]$, obs. heterogeneity
Education: Some college		0.70
		(0.68)
Education: Bachelor degree		1.91***
		(0.71)
Education: Advanced degree		2.71***
		(0.77)
Ethnicity: Black		1.82
		(2.06)
Ethnicity: Native		0.17
		(1.97)
Ethnicity: Other		-1.08
		(2.13)
Ethnicity: White		0.22
		(1.79)
Household income (thousands), $\in (5, 7.5]$		0.50
		(4.26)
Household income (thousands), $\in (7.5, 10]$		-0.54
		(2.45)
Household income (thousands), $\in (10, 12.5]$		-3.08
		(3.15)
Household income (thousands), $\in (12.5, 15]$		-0.41
		(2.44)
Household income (thousands), $\in (15, 20]$		0.27
		(2.46)
Household income (thousands), $\in (20, 25]$		-0.21
		(2.44)
Household income (thousands), $\in (25, 30]$		-1.26
		(2.37)
Household income (thousands), $\in (30, 35]$		-0.62
		(2.29)
Household income (thousands), $\in (35, 40]$		1.69
		(2.30)
Household income (thousands), $\in (40, 50]$		-0.50
		(2.21)
Household income (thousands), $\in (50, 60]$		-0.38
		(2.24)
Household income (thousands), $\in (60, 75]$		-0.30
		(2.23)
Household income (thousands), > 75 (higher cat n/a)		-1.67
		(2.24)
Household income (thousands), $\in (75, 100]$		0.22

Continued on next page

Table 1.C.1. Model with no heterogeneity and observed heterogeneity

	$E[R]$, no heterogeneity	$E[R]$, obs. heterogeneity
		(2.18)
Household income (thousands), $\in (100, 125]$		-0.77
		(2.21)
Household income (thousands), $\in (125, 200]$		-0.21
		(2.25)
Household income (thousands), $> 200k$		-1.93
		(2.39)
Ever owned stocks		-0.09
		(0.61)
Follows stock market		1.18**
		(0.56)
Understands stock market		0.07
		(0.52)
Knowledge of past returns: Don't know		1.24*
		(0.64)
Knowledge of past returns: Magnitude too large		5.07***
		(1.06)
Knowledge of past returns: Sign and Magnitude correct		3.22***
		(0.66)
Probability Numeracy		0.98***
		(0.28)
Financial Knowledge		0.29
		(0.26)
Financial Numeracy		0.19
		(0.26)
N · T	77310	32170
R^2	0.00	0.12

N = 3030 individuals for the model of the first column. Requiring all covariates drops the number of individuals to N = 806 in the second column. OLS estimates. Standard errors (clustered by individual) in parentheses. ***, ** and * denote significance at 1%, 5% and 10% respectively.

1.C.2 Sample with at least three observations per individuals

This section shows the results when we include individuals with at least 2 observations into the analysis, which is the minimum needed to calculate all the variables by which we classify individuals into groups.

Table 1.C.2. Group sizes

	N	Share of sample
group		
0	822	0.25
1	417	0.13
2	778	0.23
3	424	0.13
4	887	0.27
Total	3328	

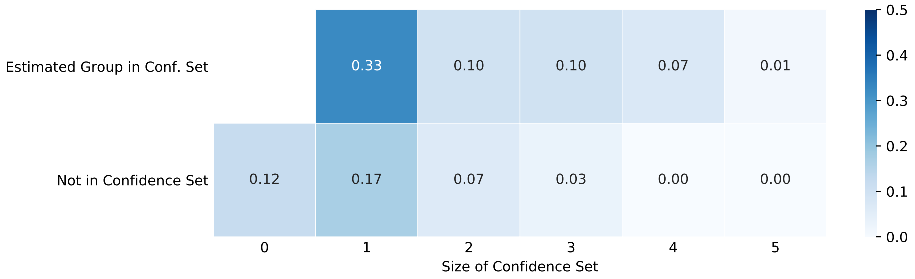


Figure 1.C.1. unit-wise 90% confidence sets by size and inclusion of estimated group

Table 1.C.3. Comparison of estimated groups

Group from alternative specification	0	1	2	3	4
Group from main specification					
Pessimists	0.97	0.00	0.01	0.00	0.02
Mean Reverters	0.05	0.00	0.90	0.00	0.05
Extrapolators	0.04	0.65	0.21	0.03	0.08
Ignorants	0.00	0.00	0.08	0.91	0.00
Sophisticates	0.00	0.00	0.02	0.00	0.98

Notes: Each row shows how individuals assigned to a given group in our main specification are allocated across groups for a different specification.

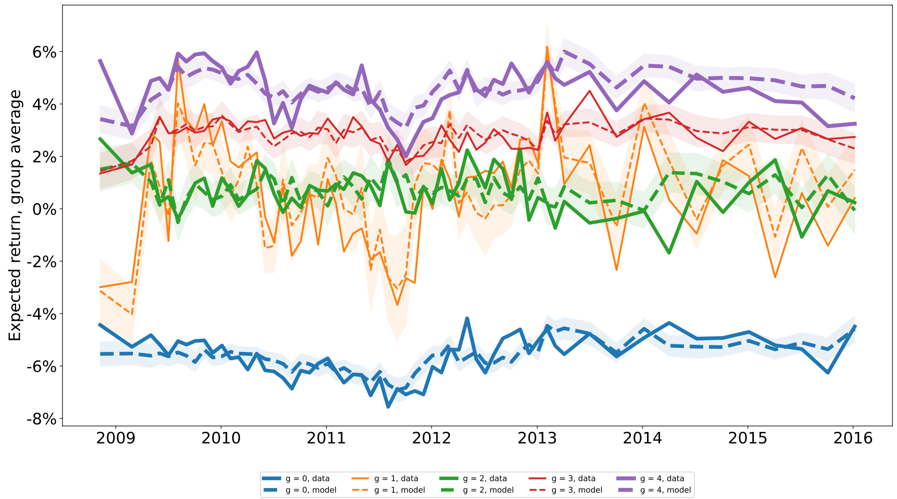
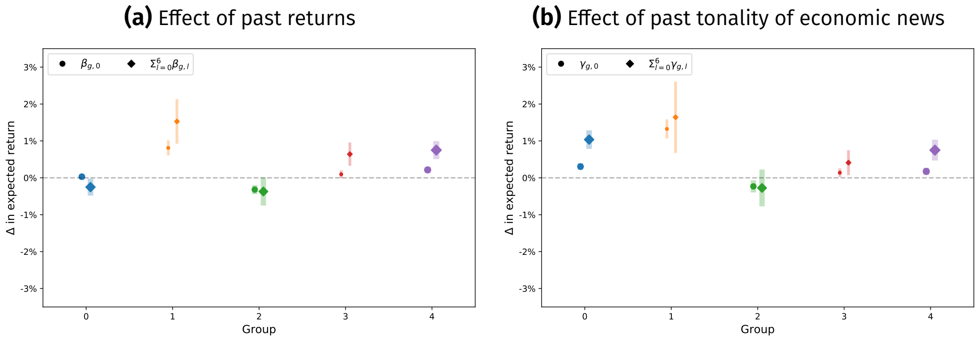


Figure 1.C.2. Data vs. predicted expected return of the Dow Jones index, by unobserved group

Notes: The solid and dashed lines are within survey and group means of individual data points and model predictions. Shaded regions are within survey and group means of individual 95% confidence intervals for the estimated regression function. Line widths are proportional to group sizes. Where within survey and group means consist of less than 15 observations, we do not plot the series, resulting in a gap. Some ALP surveys had a smaller number of individuals taking part.


(a) Effect of past returns

(b) Effect of past tonality of economic news

Figure 1.C.3. Effect on expected returns of increases in past returns and tonality of economic news, by group

Notes: Dots depict the effect on expected returns of a one standard deviation increase in the most recent monthly return of the Dow Jones (Panel a) and in the most recent tonality of economic news over one month (Panel b). Diamonds depict the summed effect in the most recent, plus six preceding monthly returns of the Dow Jones and the preceding tonalities of economic news, respectively. Shaded lines show the width of 95% confidence intervals. Marker and line widths are proportional to group sizes.

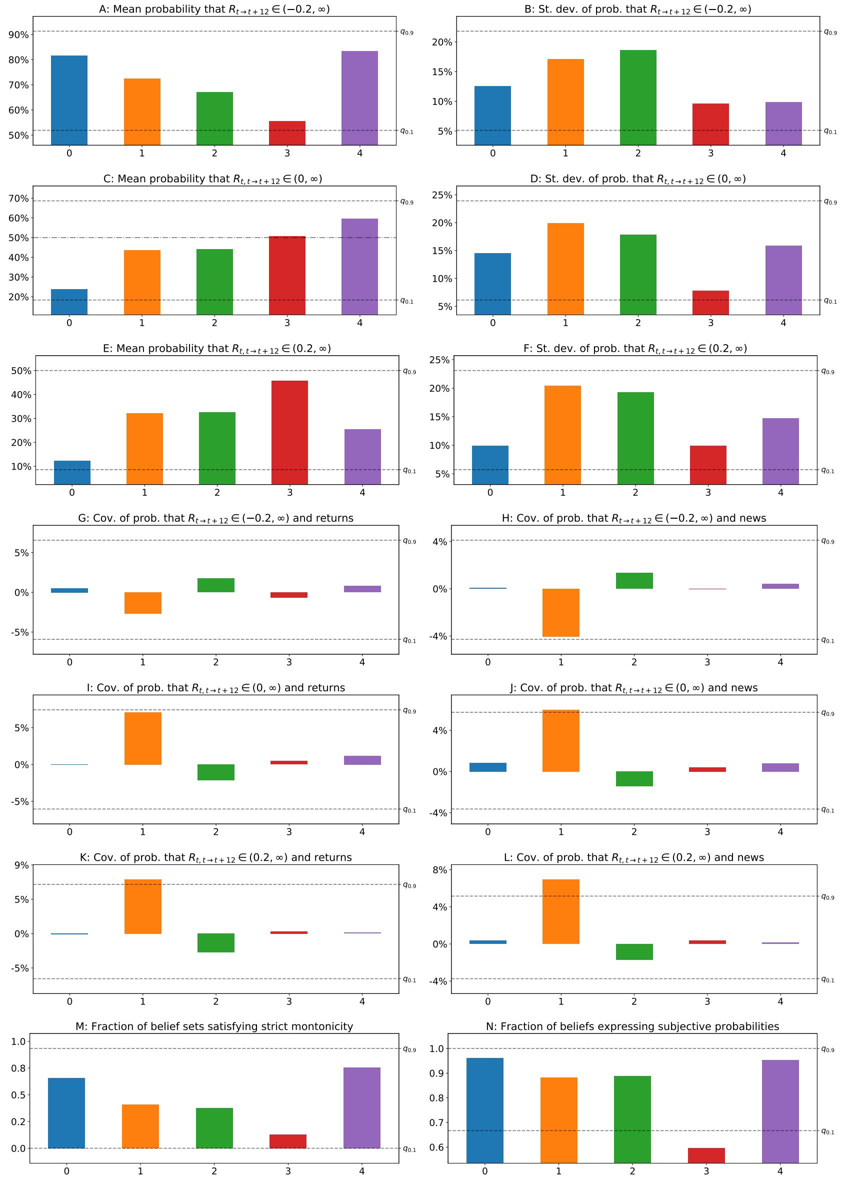


Figure 1.C.5. Moments used for classification by unobserved heterogeneity group

Notes: Bars show the group means of the 14 individual moments used to classify individuals via the k-means algorithm. Dashed lines are the bottom and top decile with respect to the individuals of all groups taken together.

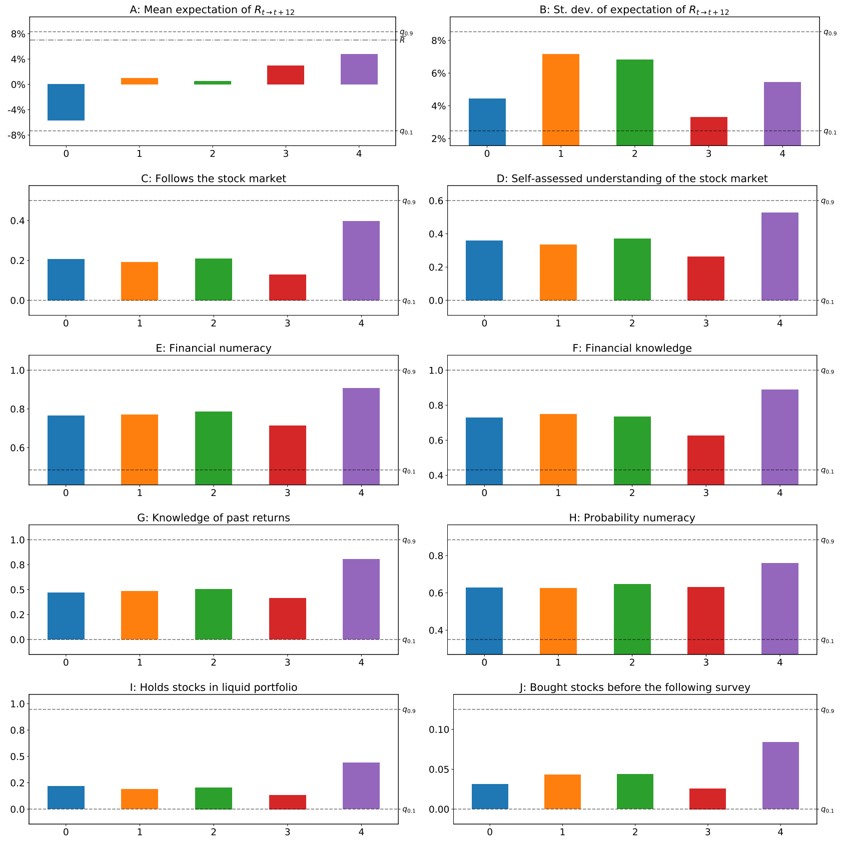


Figure 1.C.6. Characteristics not used for classification by unobserved heterogeneity group

Notes: N = 3328, smaller for some panels depending on the availability of covariates, see Table 1.1. Bars show group means, dashed lines are the bottom and top decile with respect to the individuals of all groups taken together. *Variable definitions:*: Financial numeracy and knowledge: First principle components loading on variables indicating whether a respondent correctly answered numerical and knowledge based questions, scaled to the unit interval; Probability numeracy: Fraction of correct answers to questions about probability theory; Knowledge of past returns: False sign (0), don't know (⅓), magnitude too large (⅔), sign and magnitude correct (1); Understanding of the stock market: Extremely bad (0), very bad (⅕), bad (⅖), good (⅗), very good (⅘), extremely good (1); Follows stock market: Not at all (0), somewhat (½), closely (1).

1.C.3 Sample with at least fifteen observations per individuals

This section shows the results when we include only individuals with at least 15 observations into the analysis, which is the number of parameters in our model when we include 6 lags of returns and news.

Table 1.C.4. Group sizes

group	N	Share of sample
0	475	0.23
1	367	0.18
2	410	0.20
3	264	0.13
4	526	0.26
Total	2042	

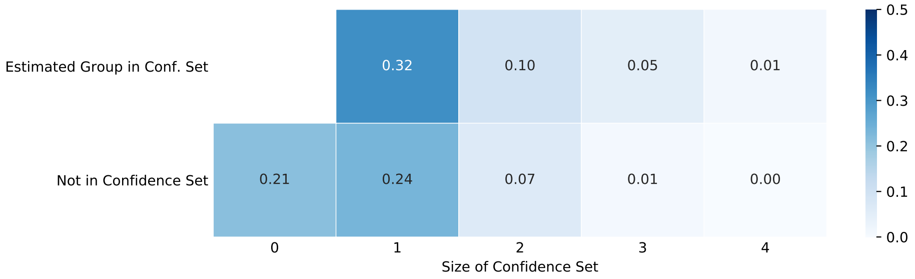


Figure 1.C.7. unit-wise 90% confidence sets by size and inclusion of estimated group

Table 1.C.5. Comparison of estimated groups

Group from alternative specification Group from main specification	0	1	2	3	4
Pessimists	0.83	0.04	0.07	0.00	0.06
Mean Reverters	0.01	0.88	0.08	0.04	0.00
Extrapolators	0.00	0.00	1.00	0.00	0.00
Ignorants	0.00	0.00	0.00	1.00	0.00
Sophisticates	0.00	0.07	0.12	0.02	0.80

Notes: Each row shows how individuals assigned to a given group in our main specification are allocated across groups for a different specification.

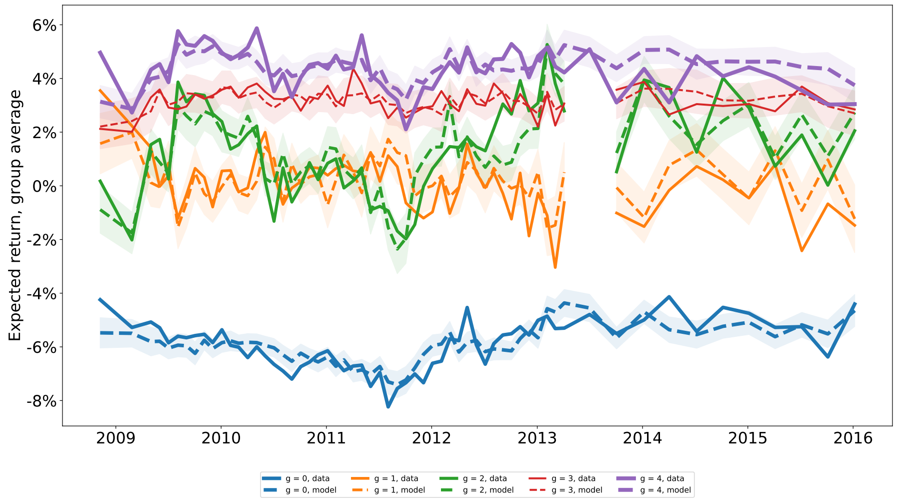
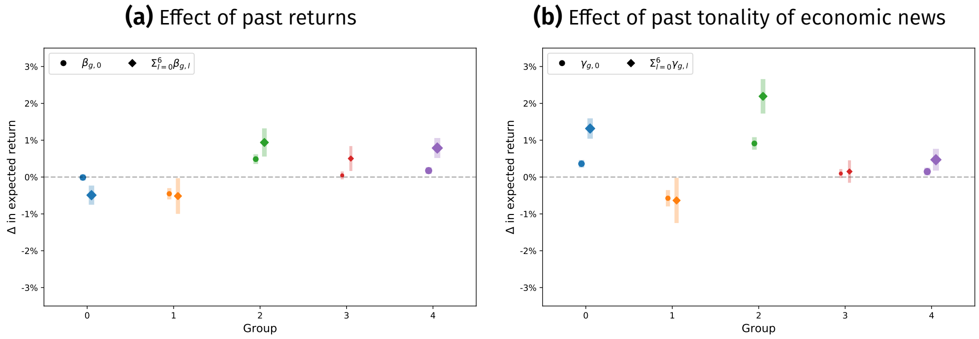


Figure 1.C.8. Data vs. predicted expected return of the Dow Jones index, by unobserved group

Notes: The solid and dashed lines are within survey and group means of individual data points and model predictions. Shaded regions are within survey and group means of individual 95% confidence intervals for the estimated regression function. Line widths are proportional to group sizes. Where within survey and group means consist of less than 15 observations, we do not plot the series, resulting in a gap. Some ALP surveys had a smaller number of individuals taking part.


(a) Effect of past returns

(b) Effect of past tonality of economic news

Figure 1.C.9. Effect on expected returns of increases in past returns and tonality of economic news, by group

Notes: Dots depict the effect on expected returns of a one standard deviation increase in the most recent monthly return of the Dow Jones (Panel a) and in the most recent tonality of economic news over one month (Panel b). Diamonds depict the summed effect in the most recent, plus six preceding monthly returns of the Dow Jones and the preceding tonalities of economic news, respectively. Shaded lines show the width of 95% confidence intervals. Marker and line widths are proportional to group sizes.

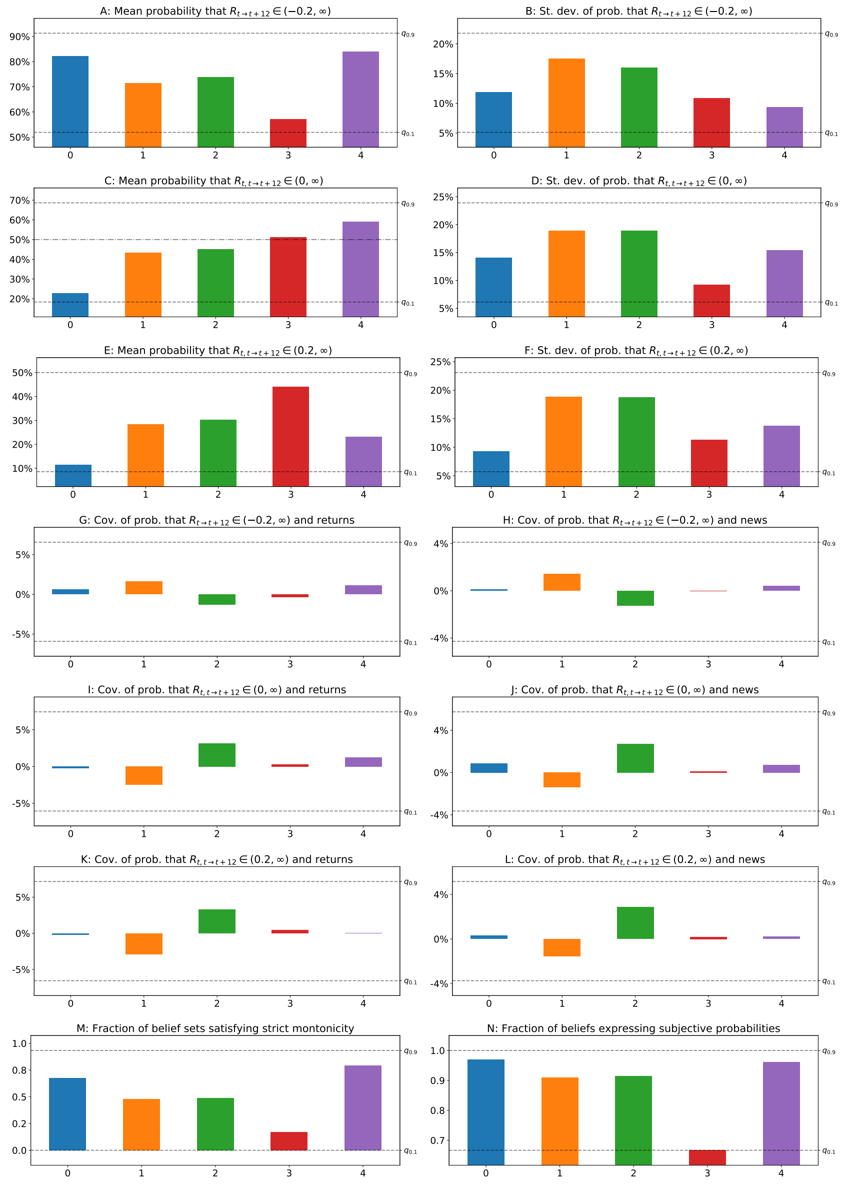


Figure 1.C.11. Moments used for classification by unobserved heterogeneity group

Notes: Bars show the group means of the 14 individual moments used to classify individuals via the k-means algorithm. Dashed lines are the bottom and top decile with respect to the individuals of all groups taken together.

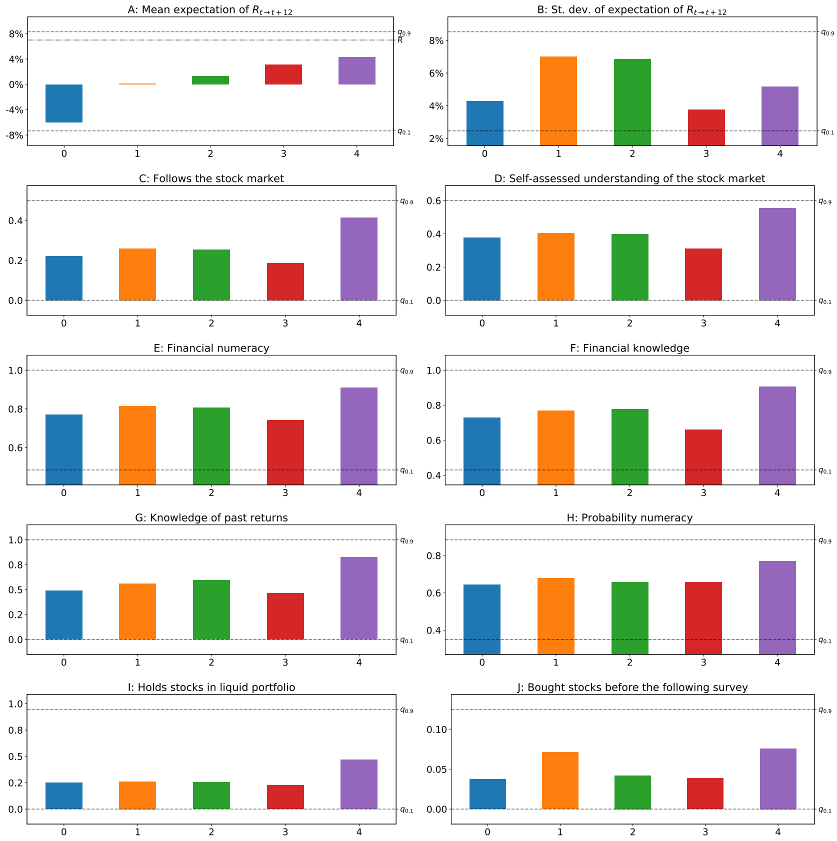


Figure 1.C.12. Characteristics not used for classification by unobserved heterogeneity group

Notes: N = 2042, smaller for some panels depending on the availability of covariates, see Table 1.1. Bars show group means, dashed lines are the bottom and top decile with respect to the individuals of all groups taken together. *Variable definitions:*: Financial numeracy and knowledge: First principle components loading on variables indicating whether a respondent correctly answered numerical and knowledge based questions, scaled to the unit interval; Probability numeracy: Fraction of correct answers to questions about probability theory; Knowledge of past returns: False sign (0), don't know (⅓), magnitude too large (⅔), sign and magnitude correct (1); Understanding of the stock market: Extremely bad (0), very bad (⅕), bad (⅖), good (⅗), very good (⅘), extremely good (1); Follows stock market: Not at all (0), somewhat (½), closely (1).

1.C.4 $\Pr(R_{t\to t+12} > 0)$ as the dependent variable

This section shows the results when we replace our dependent variable from the main analysis, $E[R_{t\to t+12}]_{i,t}$ with $\Pr(R_{t\to t+12} > 0)_{i,t}$. This substantially reduces the amount of information we use on individual beliefs, but is robust to monotonicity violations that arise when we approximate expectations using all three subjective probabilities.

Table 1.C.6. Group sizes

	N	Share of sample
group		
0	749	0.25
1	584	0.19
2	527	0.17
3	393	0.13
4	777	0.26
Total	3030	

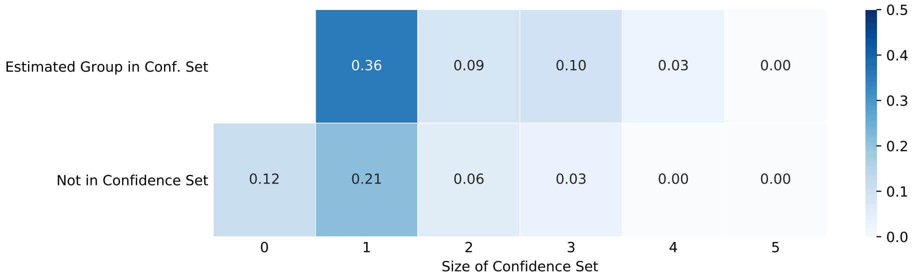


Figure 1.C.13. unit-wise 90% confidence sets by size and inclusion of estimated group

Notes: Number in cells refer to its share of the individuals.

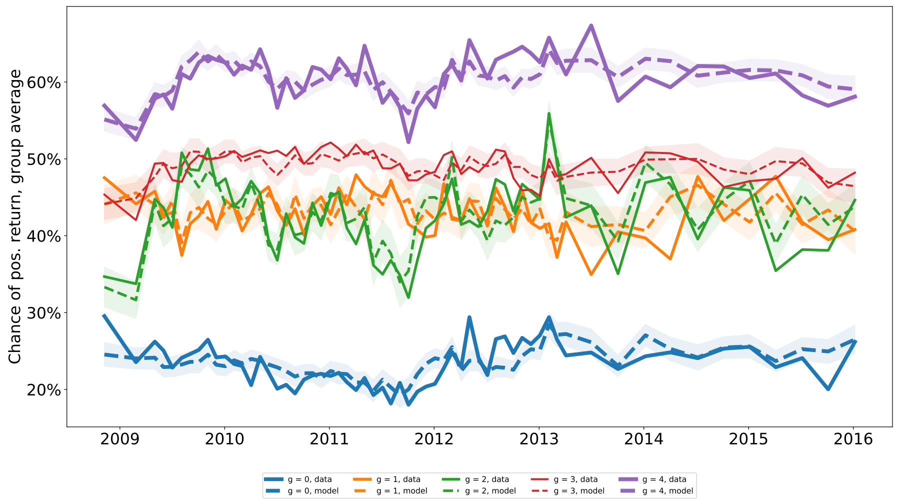
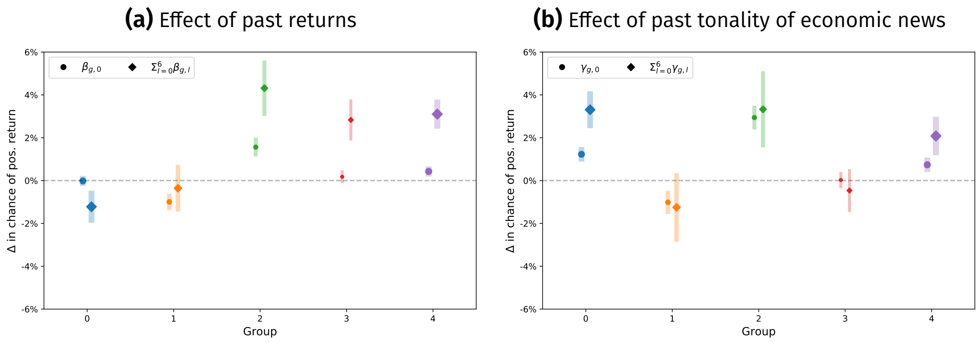


Figure 1.C.14. Data vs. predicted probability that the Dow Jones Index increases, by unobserved group

Notes: The solid and dashed lines are within survey and group means of individual data points and model predictions. Shaded regions are within survey and group means of individual 95% confidence intervals for the estimated regression function. Line widths are proportional to group sizes.


(a) Effect of past returns

(b) Effect of past tonality of economic news

Figure 1.C.15. Effect on the probability that the Dow Jones will go up of increases in past returns and tonality of economic news, by group

Notes: Dots depict the effect on expected returns of a one standard deviation increase in the most recent monthly return of the Dow Jones (Panel a) and in the most recent tonality of economic news over one month (Panel b). Diamonds depict the summed effect in the most recent, plus six preceding monthly returns of the Dow Jones and the preceding tonalities of economic news, respectively. Shaded lines show the width of 95% confidence intervals. Marker and line widths are proportional to group sizes.

1.C.5 Comparison with Dominitz-Manski types

The following plots show how our groups relate to the types of Dominitz and Manski, extended to a much longer panel, by considering the cross-sectional distribution of individual-level fractions of observations close to their specified belief types Random Walk (RW), Persistence (P) and Mean Reversion (MR).

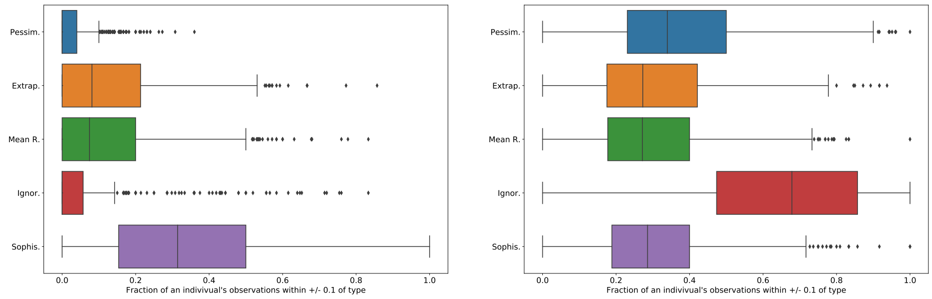


Figure 1.C.17. Cross-sectional distribution of individual compatibility with RW type, by group

Notes: Type def. in terms of response to returns

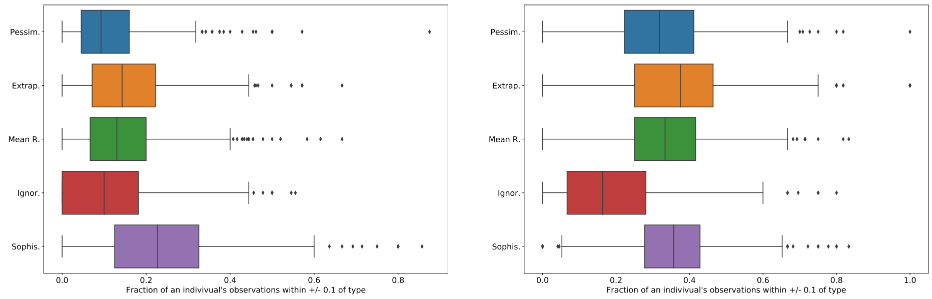


Figure 1.C.18. Cross-sectional distribution of individual compatibility with P type, by group

Notes: Type def. in terms of response to returns

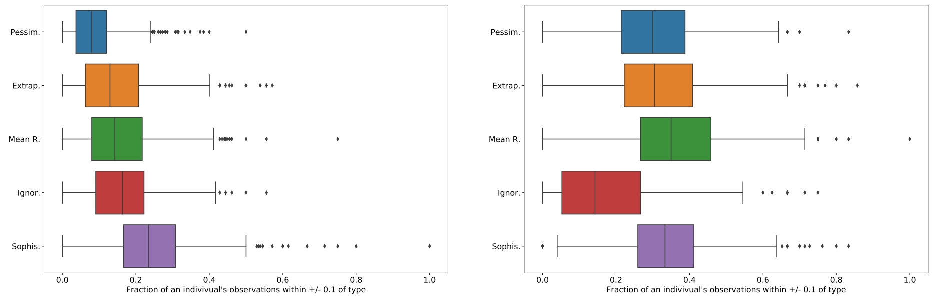


Figure 1.C.19. Cross-sectional distribution of individual compatibility with MR type, by group

Notes: Type def. in terms of response to returns

1.C.6 Three unobserved groups

This section shows the results when we assign individuals to 3 groups.

Table 1.C.7. Group sizes

group	N	Share of sample
0	1357	0.45
1	1298	0.43
2	375	0.12
Total	3030	

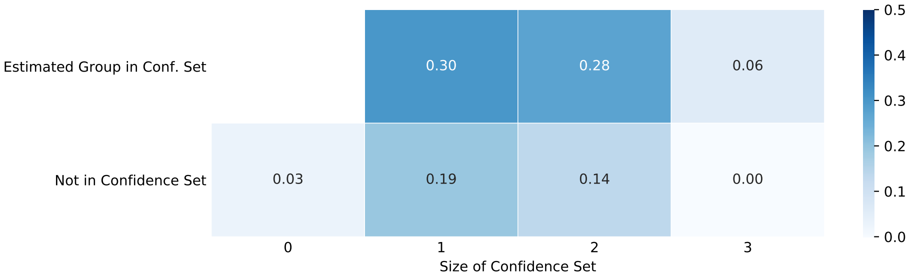


Figure 1.C.20. unit-wise 90% confidence sets by size and inclusion of estimated group

Notes: Number in cells refer to its share of the individuals.

Table 1.C.8. Comparison of estimated groups

Group from alternative specification Group from main specification	0	1	2
Pessimists	0.93	0.07	0.00
Mean Reverters	0.15	0.82	0.03
Extrapolators	0.03	0.96	0.01
Ignorants	0.00	0.12	0.88
Sophisticates	0.71	0.28	0.01

Notes: Each row shows how individuals assigned to a given group in our main specification are allocated across groups for a different specification.

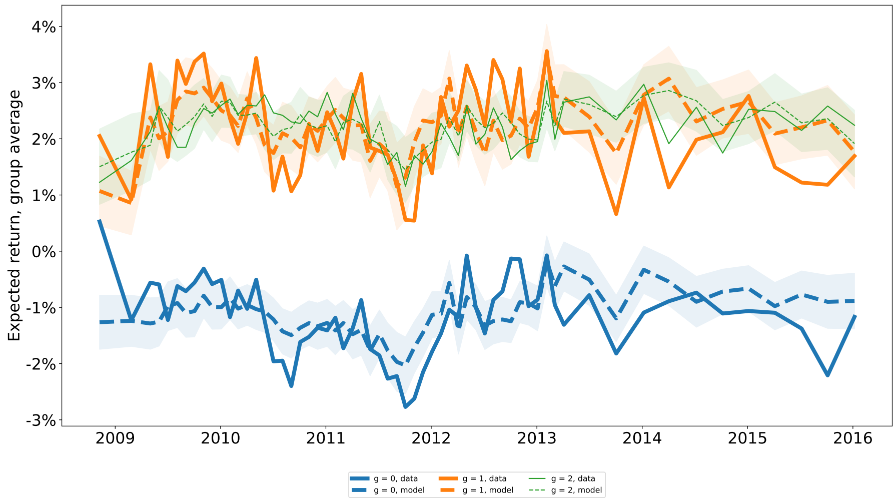
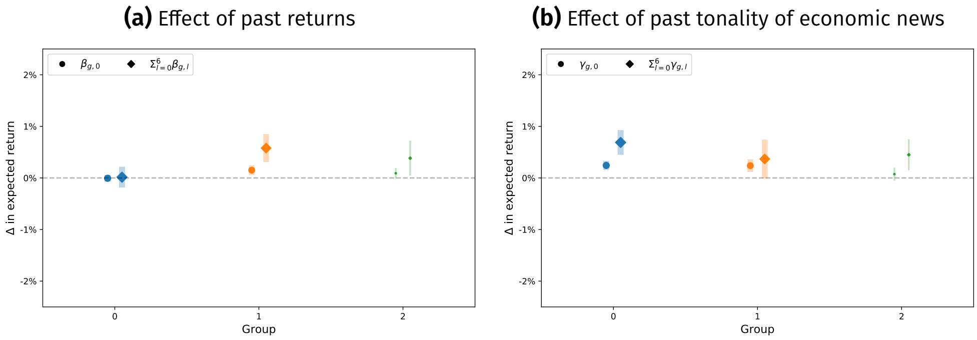


Figure 1.C.21. Data vs. predicted expected return of the Dow Jones index, by unobserved group

Notes: The solid and dashed lines are within survey and group means of individual data points and model predictions. Shaded regions are within survey and group means of individual 95% confidence intervals for the estimated regression function. Line widths are proportional to group sizes.


(a) Effect of past returns

(b) Effect of past tonality of economic news

Figure 1.C.22. Effect on expected returns of increases in past returns and tonality of economic news, by group

Notes: Dots depict the effect on expected returns of a one standard deviation increase in the most recent monthly return of the Dow Jones (Panel a) and in the most recent tonality of economic news over one month (Panel b). Diamonds depict the summed effect in the most recent, plus six preceding monthly returns of the Dow Jones and the preceding tonalities of economic news, respectively. Shaded lines show the width of 95% confidence intervals. Marker and line widths are proportional to group sizes.

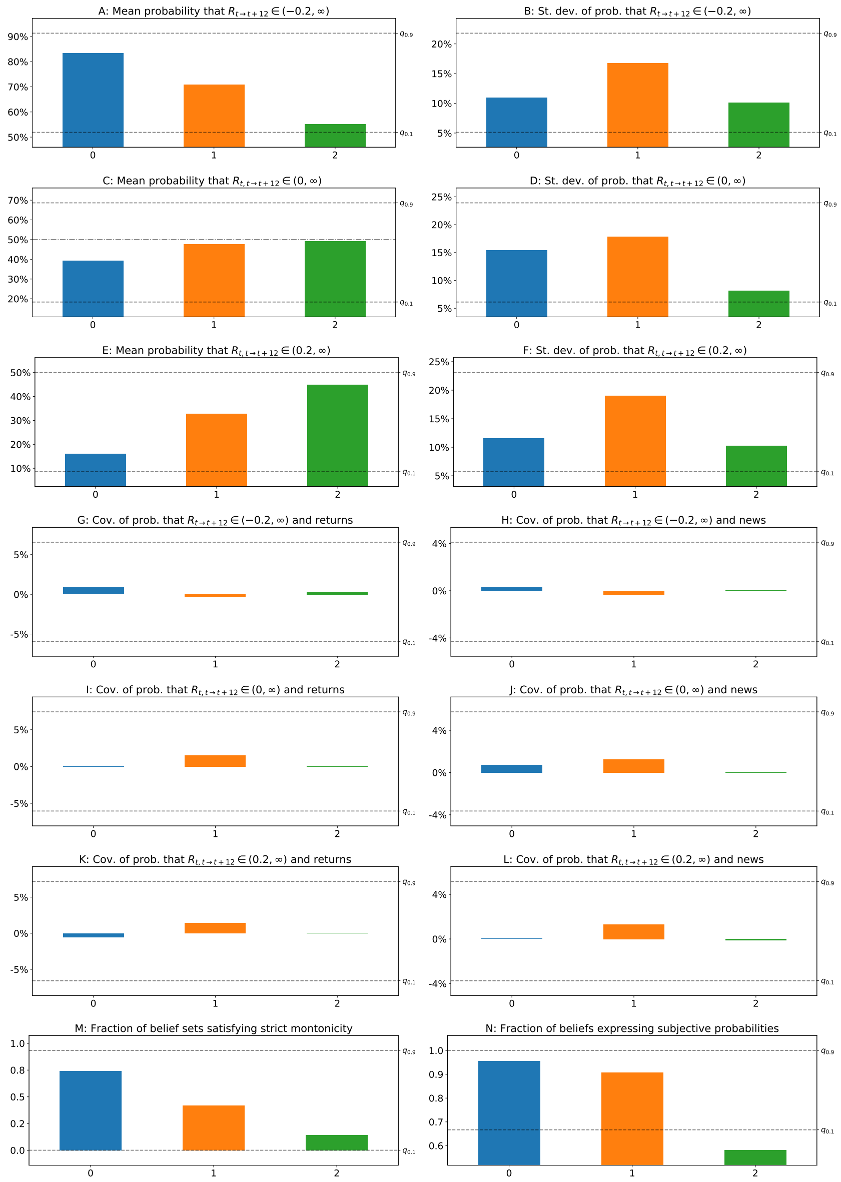


Figure 1.C.24. Moments used for classification by unobserved heterogeneity group

Notes: Bars show the group means of the 14 individual moments used to classify individuals via the k-means algorithm. Dashed lines are the bottom and top decile with respect to the individuals of all groups taken together.

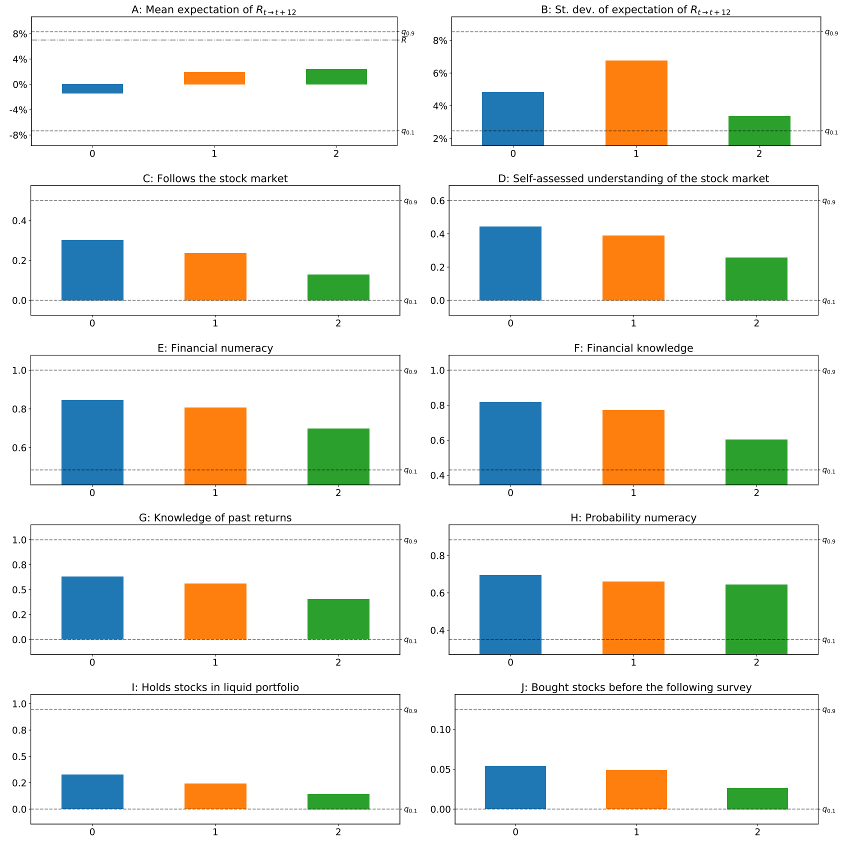


Figure 1.C.25. Characteristics not used for classification by unobserved heterogeneity group

Notes: N = 3030, smaller for some panels depending on the availability of covariates, see Table 1.1. Bars show group means, dashed lines are the bottom and top decile with respect to the individuals of all groups taken together. *Variable definitions*:: Financial numeracy and knowledge: First principle components loading on variables indicating whether a respondent correctly answered numerical and knowledge based questions, scaled to the unit interval; Probability numeracy: Fraction of correct answers to questions about probability theory; Knowledge of past returns: False sign (0), don’t know (⅓), magnitude too large (⅔), sign and magnitude correct (1); Understanding of the stock market: Extremely bad (0), very bad (⅕), bad (⅖), good (⅗), very good (⅘), extremely good (1); Follows stock market: Not at all (0), somewhat (½), closely (1).

1.C.7 Four unobserved groups

This section shows the results when we assign individuals to 4 groups.

Table 1.C.9. Group sizes

group	N	Share of sample
0	1229	0.41
1	599	0.20
2	810	0.27
3	392	0.13
Total	3030	

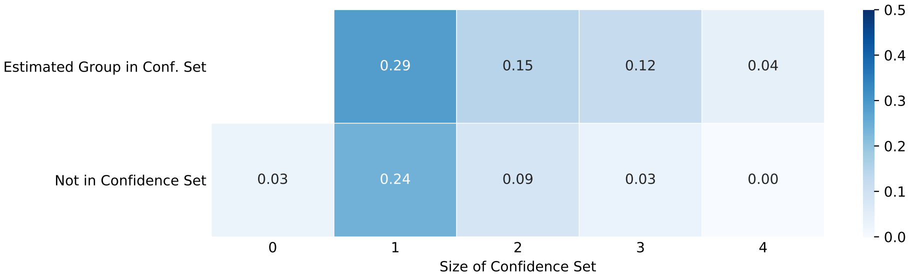


Figure 1.C.26. unit-wise 90% confidence sets by size and inclusion of estimated group

Notes: Number in cells refer to its share of the individuals.

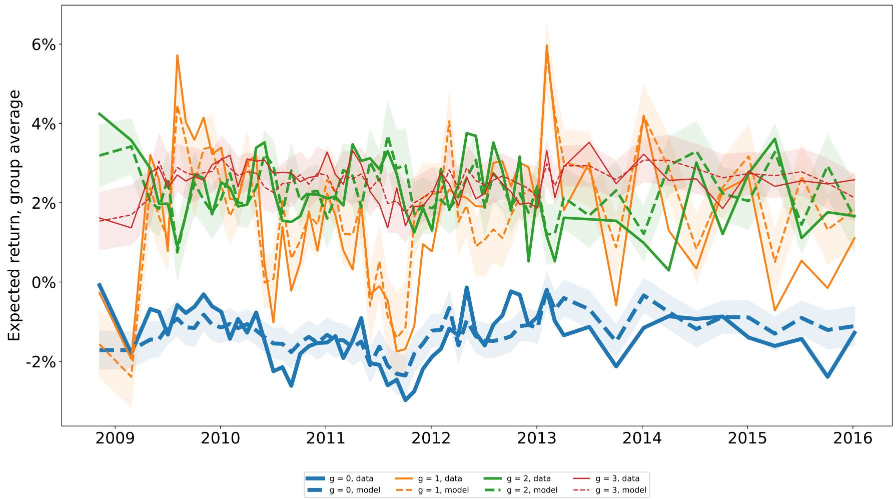
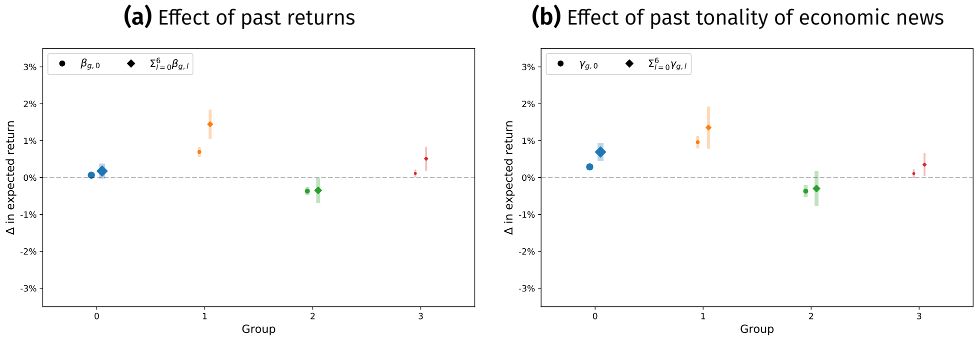


Figure 1.C.27. Data vs. predicted expected return of the Dow Jones index, by unobserved group

Notes: The solid and dashed lines are within survey and group means of individual data points and model predictions. Shaded regions are within survey and group means of individual 95% confidence intervals for the estimated regression function. Line widths are proportional to group sizes.


(a) Effect of past returns

(b) Effect of past tonality of economic news

Figure 1.C.28. Effect on expected returns of increases in past returns and tonality of economic news, by group

Notes: Dots depict the effect on expected returns of a one standard deviation increase in the most recent monthly return of the Dow Jones (Panel a) and in the most recent tonality of economic news over one month (Panel b). Diamonds depict the summed effect in the most recent, plus six preceding monthly returns of the Dow Jones and the preceding tonalities of economic news, respectively. Shaded lines show the width of 95% confidence intervals. Marker and line widths are proportional to group sizes.

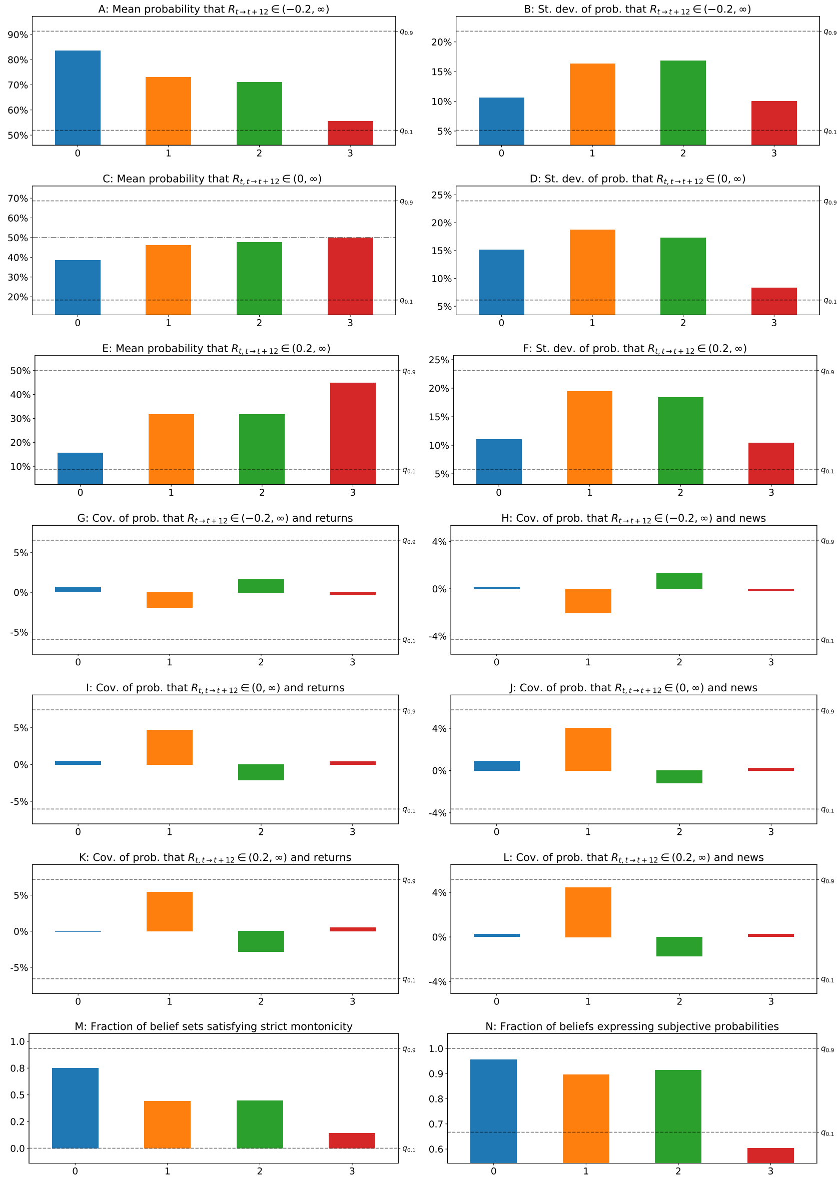


Figure 1.C.30. Moments used for classification by unobserved heterogeneity group

Notes: Bars show the group means of the 14 individual moments used to classify individuals via the k-means algorithm. Dashed lines are the bottom and top decile with respect to the individuals of all groups taken together.

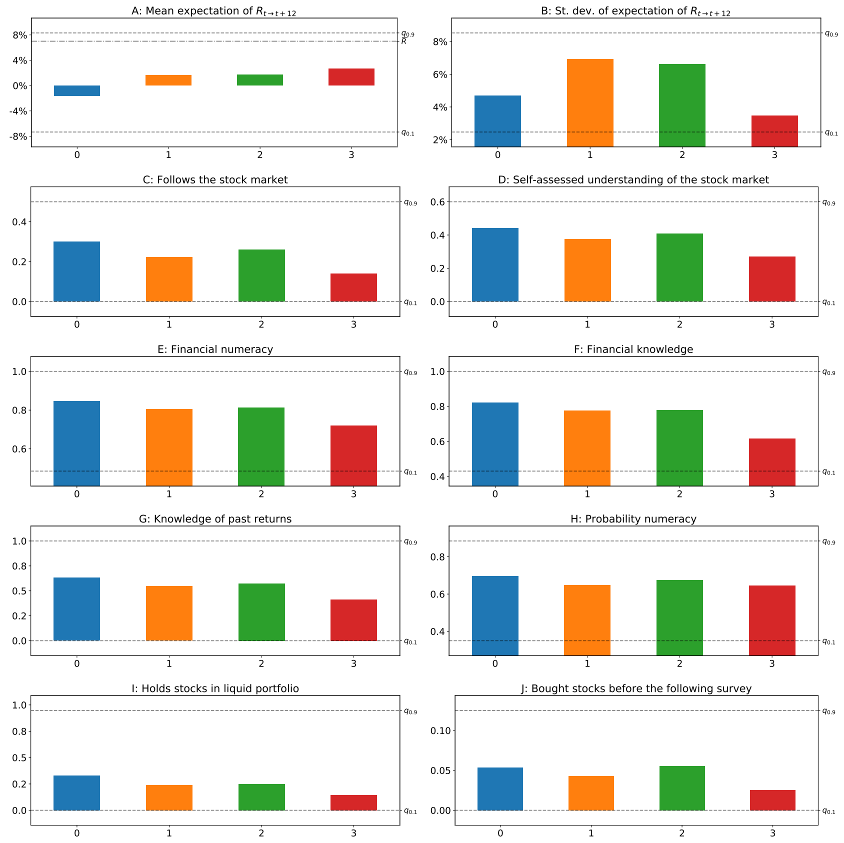


Figure 1.C.31. Characteristics not used for classification by unobserved heterogeneity group

Notes: N = 3030, smaller for some panels depending on the availability of covariates, see Table 1.1. Bars show group means, dashed lines are the bottom and top decile with respect to the individuals of all groups taken together. *Variable definitions:*: Financial numeracy and knowledge: First principle components loading on variables indicating whether a respondent correctly answered numerical and knowledge based questions, scaled to the unit interval; Probability numeracy: Fraction of correct answers to questions about probability theory; Knowledge of past returns: False sign (0), don't know (⅓), magnitude too large (⅔), sign and magnitude correct (1); Understanding of the stock market: Extremely bad (0), very bad (⅕), bad (⅖), good (⅗), very good (⅘), extremely good (1); Follows stock market: Not at all (0), somewhat (½), closely (1).

Table 1.C.10. Comparison of estimated groups

Group from alternative specification Group from main specification	0	1	2	3
Pessimists	0.94	0.01	0.05	0.00
Mean Reverters	0.01	0.00	0.99	0.00
Extrapolators	0.00	0.97	0.02	0.01
Ignorants	0.00	0.00	0.04	0.96
Sophisticates	0.67	0.10	0.22	0.01

Notes: Each row shows how individuals assigned to a given group in our main specification are allocated across groups for a different specification.

1.C.8 Seven unobserved groups

This section shows the results when we assign individuals to 7 groups.

Table 1.C.11. Group sizes

group	N	Share of sample
0	599	0.20
1	424	0.14
2	249	0.08
3	296	0.10
4	216	0.07
5	520	0.17
6	726	0.24
Total	3030	

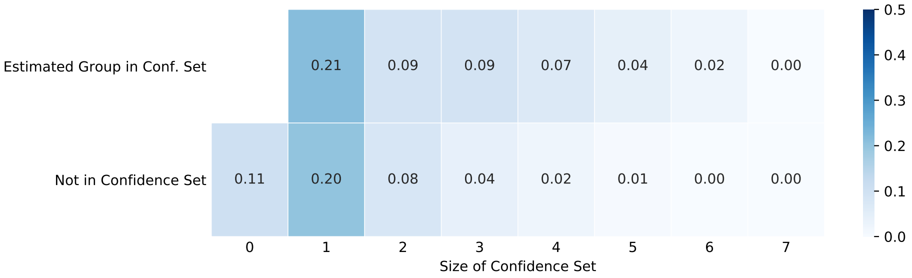


Figure 1.C.32. unit-wise 90% confidence sets by size and inclusion of estimated group

Notes: Number in cells refer to its share of the individuals.

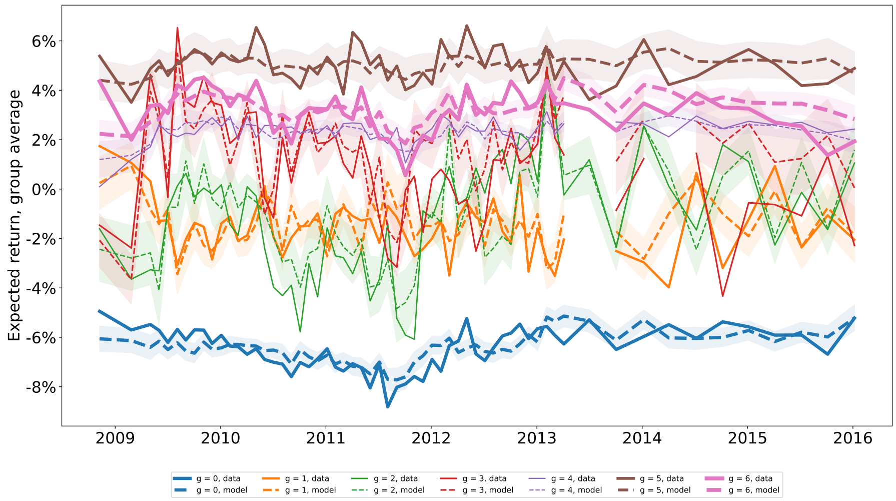
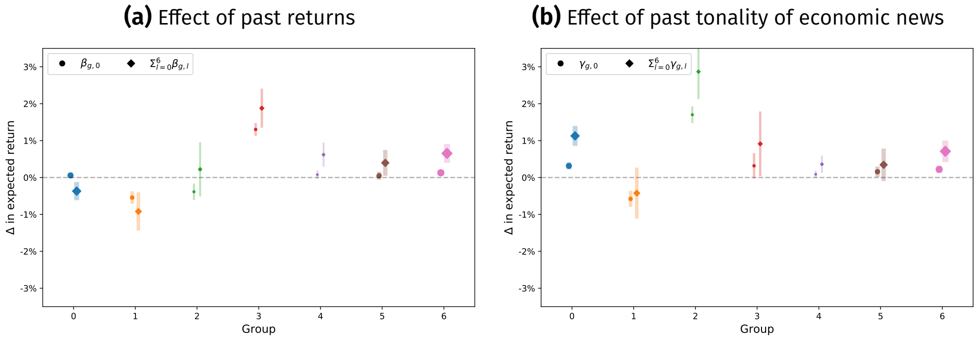


Figure 1.C.33. Data vs. predicted expected return of the Dow Jones index, by unobserved group

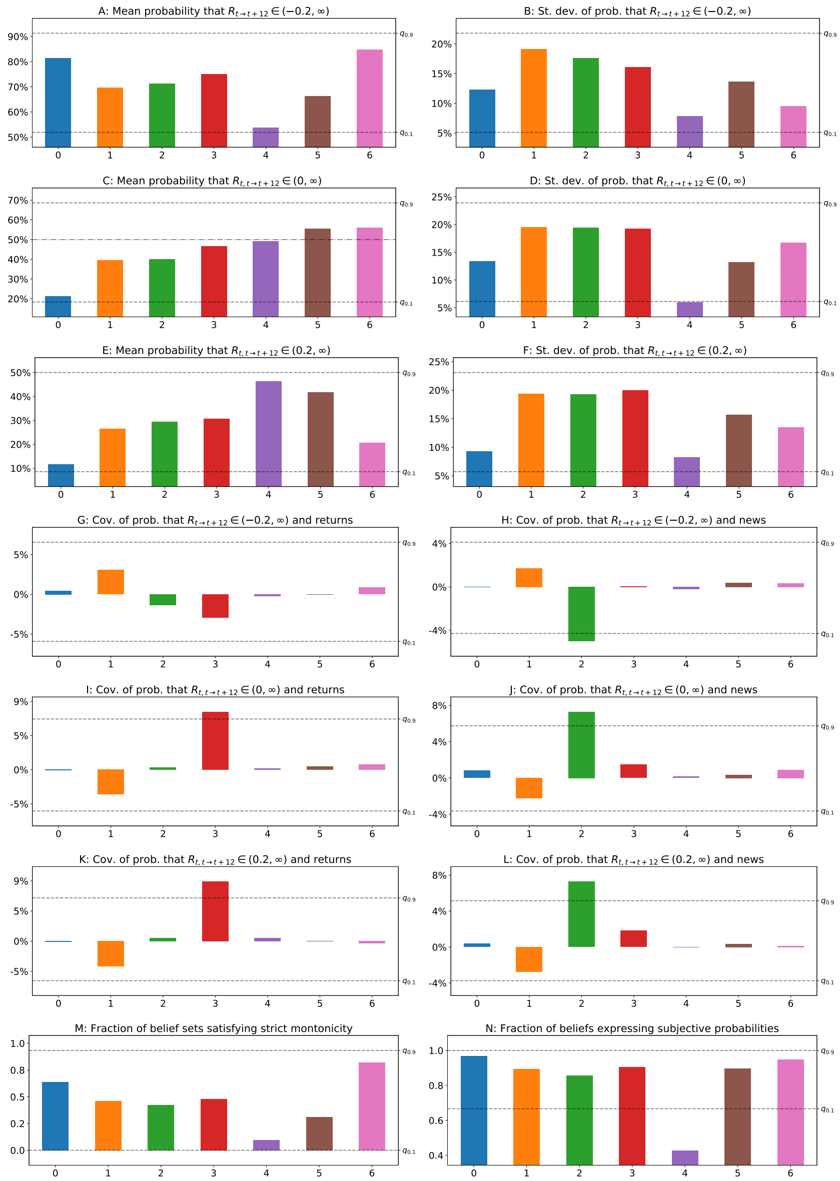
Notes: The solid and dashed lines are within survey and group means of individual data points and model predictions. Shaded regions are within survey and group means of individual 95% confidence intervals for the estimated regression function. Line widths are proportional to group sizes. Where within survey and group means consist of less than 15 observations, we do not plot the series, resulting in a gap. Some ALP surveys had a small number of individuals taking part.


(a) Effect of past returns

(b) Effect of past tonality of economic news

Figure 1.C.34. Effect on expected returns of increases in past returns and tonality of economic news, by group

Notes: Dots depict the effect on expected returns of a one standard deviation increase in the most recent monthly return of the Dow Jones (Panel a) and in the most recent tonality of economic news over one month (Panel b). Diamonds depict the summed effect in the most recent, plus six preceding monthly returns of the Dow Jones and the preceding tonalities of economic news, respectively. Shaded lines show the width of 95% confidence intervals. Marker and line widths are proportional to group sizes.


A: Mean probability that $R_{t\to t+12} \in (-0.2, \infty)$

B: St. dev. of prob. that $R_{t\to t+12} \in (-0.2, \infty)$

C: Mean probability that $R_{t,t\to t+12} \in (0, \infty)$

D: St. dev. of prob. that $R_{t,t\to t+12} \in (0, \infty)$

E: Mean probability that $R_{t,t\to t+12} \in (0.2, \infty)$

F: St. dev. of prob. that $R_{t,t\to t+12} \in (0.2, \infty)$

G: Cov. of prob. that $R_{t\to t+12} \in (-0.2, \infty)$ and returns

H: Cov. of prob. that $R_{t\to t+12} \in (-0.2, \infty)$ and news

I: Cov. of prob. that $R_{t,t\to t+12} \in (0, \infty)$ and returns

J: Cov. of prob. that $R_{t,t\to t+12} \in (0, \infty)$ and news

K: Cov. of prob. that $R_{t,t\to t+12} \in (0.2, \infty)$ and returns

L: Cov. of prob. that $R_{t,t\to t+12} \in (0.2, \infty)$ and news

M: Fraction of belief sets satisfying strict montonicity

N: Fraction of beliefs expressing subjective probabilities

Figure 1.C.36. Moments used for classification by unobserved heterogeneity group

Notes: Bars show the group means of the 14 individual moments used to classify individuals via the k-means algorithm. Dashed lines are the bottom and top decile with respect to the individuals of all groups taken together.

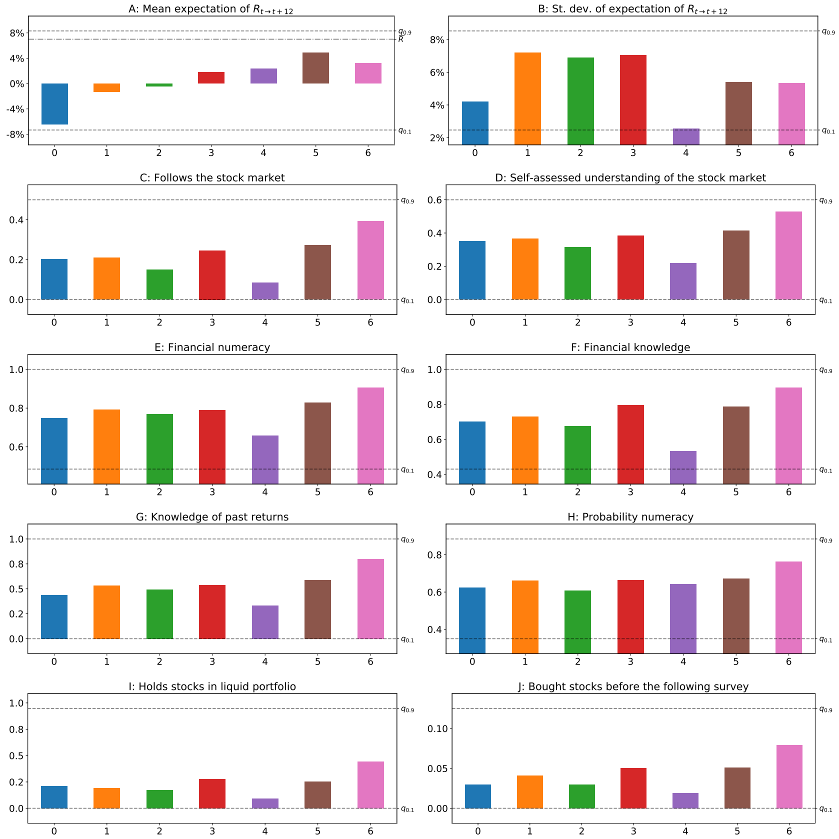


Figure 1.C.37. Characteristics not used for classification by unobserved heterogeneity group

Notes: N = 3030, smaller for some panels depending on the availability of covariates, see Table 1.1. Bars show group means, dashed lines are the bottom and top decile with respect to the individuals of all groups taken together. *Variable definitions:*: Financial numeracy and knowledge: First principle components loading on variables indicating whether a respondent correctly answered numerical and knowledge based questions, scaled to the unit interval; Probability numeracy: Fraction of correct answers to questions about probability theory; Knowledge of past returns: False sign (0), don't know (⅓), magnitude too large (⅔), sign and magnitude correct (1); Understanding of the stock market: Extremely bad (0), very bad (⅕), bad (⅖), good (⅗), very good (⅘), extremely good (1); Follows stock market: Not at all (0), somewhat (½), closely (1).

Table 1.C.12. Comparison of estimated groups

Group from alternative specification Group from main specification	0	1	2	3	4	5	6
Pessimists	0.79	0.02	0.02	0.02	0.00	0.01	0.15
Mean Reverters	0.00	0.69	0.02	0.02	0.00	0.26	0.01
Extrapolators	0.01	0.01	0.43	0.43	0.00	0.12	0.01
Ignorants	0.00	0.01	0.01	0.01	0.54	0.44	0.00
Sophisticates	0.00	0.00	0.00	0.06	0.00	0.17	0.78

Notes: Each row shows how individuals assigned to a given group in our main specification are allocated across groups for a different specification.

1.C.9 Fifteen unobserved groups

This section shows the results when we assign individuals to 15 groups.

Table 1.C.13. Group sizes

	N	Share of sample
group		
0	287	0.09
1	263	0.09
2	124	0.04
3	80	0.03
4	364	0.12
5	207	0.07
6	165	0.05
7	124	0.04
8	322	0.11
9	154	0.05
10	219	0.07
11	113	0.04
12	22	0.01
13	379	0.13
14	207	0.07
Total	3030	

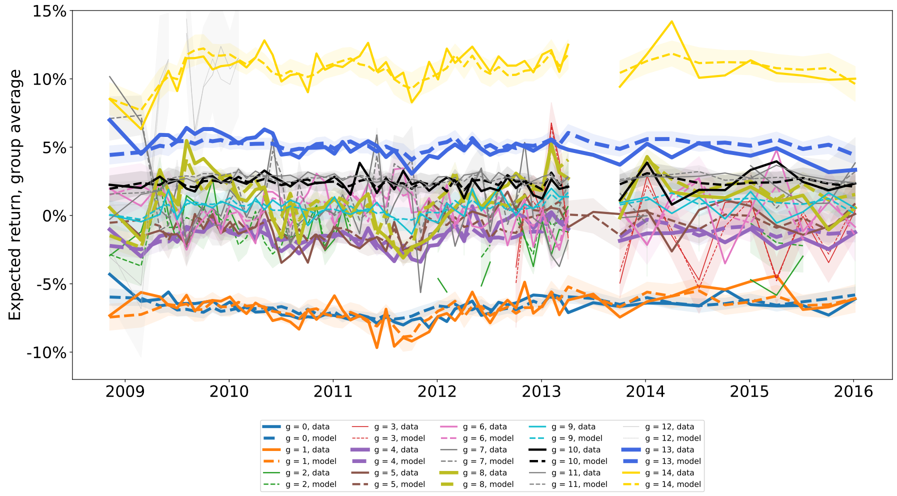
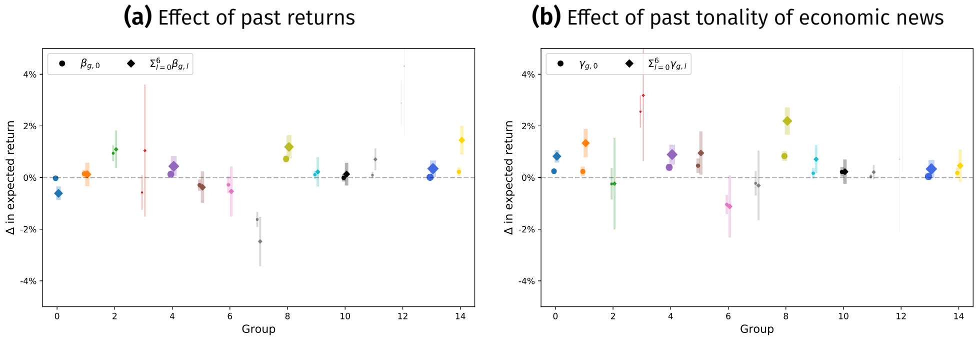


Figure 1.C.38. Data vs. predicted expected return of the Dow Jones index, by unobserved group

Notes: The solid and dashed lines are within survey and group means of individual data points and model predictions. Shaded regions are within survey and group means of individual 95% confidence intervals for the estimated regression function. Line widths are proportional to group sizes. Where within survey and group means consist of less than 15 observations, we do not plot the series, resulting in a gap. Some ALP surveys had a small number of individuals taking part.


(a) Effect of past returns

(b) Effect of past tonality of economic news

Figure 1.C.39. Effect on expected returns of increases in past returns and tonality of economic news, by group

Notes: Dots depict the effect on expected returns of a one standard deviation increase in the most recent monthly return of the Dow Jones (Panel a) and in the most recent tonality of economic news over one month (Panel b). Diamonds depict the summed effect in the most recent, plus six preceding monthly returns of the Dow Jones and the preceding tonalities of economic news, respectively. Shaded lines show the width of 95% confidence intervals. Marker and line widths are proportional to group sizes.

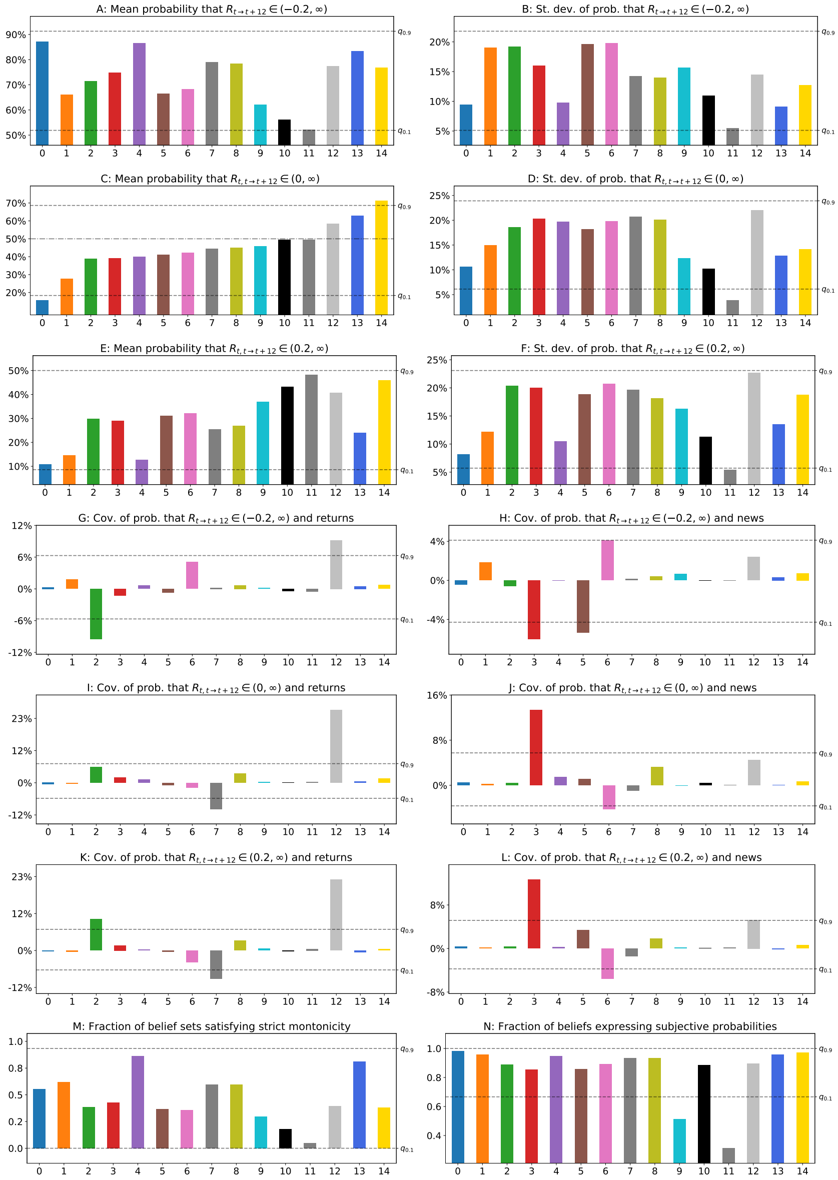


Figure 1.C.41. Moments used for classification by unobserved heterogeneity group

Notes: Bars show the group means of the 14 individual moments used to classify individuals via the k-means algorithm. Dashed lines are the bottom and top decile with respect to the individuals of all groups taken together.

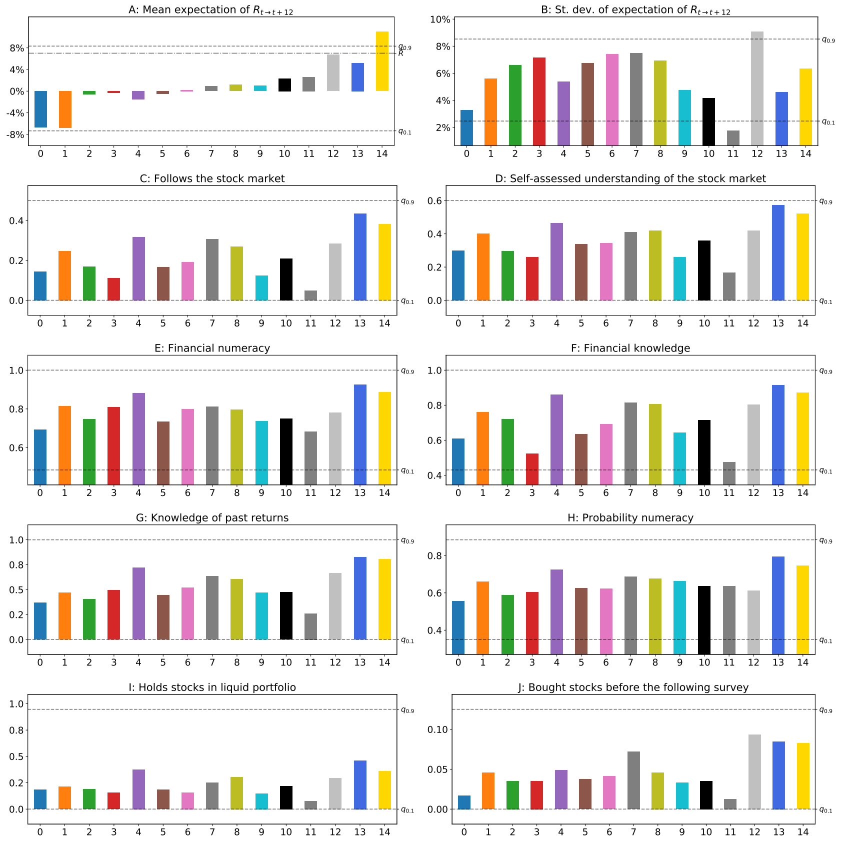


Figure 1.C.42. Characteristics not used for classification by unobserved heterogeneity group

Notes: N = 3030, smaller for some panels depending on the availability of covariates, see Table 1.1. Bars show group means, dashed lines are the bottom and top decile with respect to the individuals of all groups taken together. *Variable definitions:*: Financial numeracy and knowledge: First principle components loading on variables indicating whether a respondent correctly answered numerical and knowledge based questions, scaled to the unit interval; Probability numeracy: Fraction of correct answers to questions about probability theory; Knowledge of past returns: False sign (0), don't know (⅓), magnitude too large (⅔), sign and magnitude correct (1); Understanding of the stock market: Extremely bad (0), very bad (⅕), bad (⅖), good (⅗), very good (⅘), extremely good (1); Follows stock market: Not at all (0), somewhat (½), closely (1).

Table 1.C.14. Comparison of estimated groups

Group from alternative specification	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Group from main specification															
Pessimists	0.38	0.22	0.01	0.00	0.30	0.02	0.00	0.01	0.05	0.00	0.01	0.00	0.00	0.01	0.00
Mean Reverters	0.00	0.14	0.03	0.00	0.00	0.11	0.28	0.16	0.07	0.05	0.06	0.00	0.00	0.00	0.09
Extrapolators	0.00	0.02	0.19	0.15	0.00	0.24	0.00	0.00	0.25	0.03	0.02	0.00	0.04	0.00	0.05
Ignorants	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.26	0.41	0.28	0.00	0.00	0.04
Sophisticates	0.00	0.01	0.00	0.00	0.18	0.00	0.00	0.03	0.14	0.01	0.01	0.00	0.00	0.48	0.15

Notes: Each row shows how individuals assigned to a given group in our main specification are allocated across groups for a different specification.

Appendix 1.D Additional details on stock ownership and trading

Table 1.D.1. Regression coefficients underlying Figure 1.8

	Bought stocks in following period
Pessimists	-1.655***
	(0.066)
Mean Reverters	-1.560***
	(0.073)
Extrapolators	-1.768***
	(0.069)
Ignorants	-1.920***
	(0.083)
Sophisticates	-1.410***
	(0.047)
E[R]	0.010***
	(0.002)
N · T	70114
Pseudo R^2	0.02

Notes: N = 3029. Probit regression of the dummy indicating whether stocks were bought in the next period (if within 120 days) on expected returns and group indicators. Expected returns in percentage points. Standard errors clustered at individual level.

Table 1.D.2. Stock buying v expected returns

	Average marginal effect	Std. err.
Pessimists	0.09	0.03
Mean Reverters	0.12	0.03
Extrapolators	0.09	0.02
Ignorants	0.07	0.02
Sophisticates	0.16	0.04

Notes: N = 3029. From a probit regression of the dummy indicating whether stocks were bought in the next period (if within 120 days) on expected returns and group indicators. Both units in percentage points. Standard errors clustered at individual level. Average marginal effects calculated within group.

Appendix 1.E Additional details for forecast error analysis

Table 1.E.1 shows the same analysis as Table 1.6 in the main text, but uses our entire sample as the basis.

Table 1.E.1. Predictability of forecast errors with forecast revisions, full sample

	Pooled OLS	Pooled OLS w groups
Forecast Revision	-0.51	
	(0.02)	
Forecast Revision, Pessimists		-0.52
		(0.02)
Forecast Revision, Mean Reverters		-0.50
		(0.01)
Forecast Revision, Extrapolators		-0.50
		(0.03)
Forecast Revision, Ignorants		-0.46
		(0.03)
Forecast Revision, Sophisticates		-0.53
		(0.04)
R^2	0.10	0.21
N · T	74165	74165

Notes: N = 3030. OLS estimates. Standard errors (clustered by individual and survey) in parentheses.

References

Adam, Klaus, Albert Marcet, and Johannes Beutel. 2017. "Stock price booms and expected capital gains." *American Economic Review* 107: 2352–408. [6]

Adam, Klaus, Albert Marcet, and Juan Pablo Nicolini. 1, 2016. "Stock Market Volatility and Learning." *The Journal of Finance* 71 (1): 33–82. [6]

Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer. 2015. "X-CAPM: An Extrapolative Capital Asset Pricing Model." *Journal of Financial Economics* 115: 1–24. [6, 30]

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa. 2017. "Discretizing Unobserved Heterogeneity." IFS Working Paper W17/03. [6, 15, 16]

Bonhomme, Stéphane, and Elena Manresa. 2015. "Grouped patterns of heterogeneity in panel data." *Econometrica* 83 (3): 1147–84. [18]

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer. 2018. "Over-reaction in macroeconomic expectations." Working paper. National Bureau of Economic Research. [7, 27]

Bruin, Wändi Bruine de, Baruch Fischhoff, Susan G. Millstein, and Bonnie L. Halpern-Felsher. 2000. "Verbal and Numerical Expressions of Probability: "It's a Fifty–Fifty Chance"." *Organizational Behavior and Human Decision Processes* 81 (1): 115–31. [12, 42]

Campbell, John Y, and Samuel B Thompson. 2007. "Predicting excess stock returns out of sample: Can anything beat the historical average?" *Review of Financial Studies* 21 (4): 1509–31. [27]

Coibion, Olivier, and Yuriy Gorodnichenko. 2015. "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts." *American Economic Review* 105 (8): 2644–78. [7, 26, 27]

Dominitz, Jeff, and Charles F. Manski. 2007. "Expected Equity Returns and Portfolio Choice: Evidence from the Health and Retirement Study." *Journal of the European Economic Association* 5 (2/3): 369–79. [11]

Dominitz, Jeff, and Charles F. Manski. 2011. "Measuring and Interpreting Expectations of Equity Returns." *Journal of Applied Econometrics* 26 (3): 352–70. [6, 25, 29]

Drerup, Tilman, Benjamin Enke, and Hans-Martin von Gaudecker. 2017. "The Precision of Subjective Data and the Explanatory Power of Economic Models." *Journal of Econometrics*. Measurement Error Models 200 (2): 378–89. [41]

Dzemski, Andreas, and Ryo Okui. 2018. "Confidence Set for Group Membership." Available at https://adzemski.github.io/research/. [5, 6, 17, 29]

Enke, Benjamin, and Thomas Graeber. 2019. "Cognitive Uncertainty." Working Paper. [12]

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen P. Utkus. 2019. "Five Facts About Beliefs and Portfolios." Available at: http://dx.doi.org/10.2139/ssrn.3336400. [7, 23, 26, 29]

Gouret, Fabian, and Guillaume Hollard. 2011. "When Kahneman Meets Manski: Using Dual Systems of Reasoning to Interpret Subjective Expectations of Equity Returns." *Journal of Applied Econometrics* 26 (3): 371–92. [44]

Greenwood, Robin, and Andrei Shleifer. 2014. "Expectations of Returns and Expected Returns." *Review of Financial Studies* 27 (3): 714–46. [5, 12, 14, 26]

Hudomiet, Péter, Michael D. Hurd, and Susann Rohwedder. 2018. "Measuring Probability Numeracy." RAND Labor and Population Working Paper, WR-1270. [9, 10]

Hudomiet, Péter, Gabor Kezdi, and Robert J. Willis. 2010. "Stock market crash and expectations of American households." *Journal of Applied Econometrics*, [43]

Hurd, Michael D. 2009. "Subjective Probabilities in Household Surveys." *Annual Review of Economics* 1 (1): 543–62. [5, 12, 37]

Hurd, Michael D., and Susann Rohwedder. 2011. "Stock Price Expectations and Stock Trading." RAND Labor and Population Working Paper 938. [6, 8, 10]

Hurd, Michael D., Maarten C. J. van Rooij, and Joachim Winter. 2011. "Stock Market Expectations of Dutch Households." *Journal of Applied Econometrics* 26 (3): 416–36. [11, 37, 41, 43]

Kezdi, Gabor, and Robert J. Willis. 2008. "Stock market expectations and portfolio choice of American households." [12, 44]

Kleinjans, Kristin J., and Arthur Van Soest. 2014. "Rounding, Focal Point Answers and Nonresponse to Subjective Probability Questions." *Journal of Applied Econometrics* 29 (4): 567–85. [41]

Lusardi, Annamaria, and Olivia S. Mitchell. 2007. "Financial Literacy and Retirement Planning: New Evidence from the RAND American Life Panel." University of Pennsylvania, The Wharton School, Pension Research Council Working Paper WP2007-33. [9, 10]

Lusardi, Annamaria, and Olivia S. Mitchell. 2014. "The Economic Importance of Financial Literacy: Theory and Evidence." *Journal of Economic Literature* 52 (1): 5–44. [8]

Malmendier, Ulrike, and Stefan Nagel. 2011. "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?" *Quarterly Journal of Economics* 126 (1): 373–416. [5]

Manski, Charles F. 2004. "Measuring Expectations." *Econometrica* 72 (5): 1329–76. [5]

Manski, Charles F., and Francesca Molinari. 2010. "Rounding probabilistic expectations in surveys." *Journal of Business and Economic Statistics* 28: 219–31. [41]

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–30. [15]

Chapter 2

The distribution of ambiguity attitudes*

Joint with Hans-Martin von Gaudecker and Christian Zimpelmann

2.1 Introduction

Economists typically assume that households have precise knowledge of the relevant probability distribution when taking decisions in non-deterministic contexts. There is mounting evidence that this may not be the case. Elicitations of subjective beliefs regularly reveal violations of the basic axioms of probability theory (e.g. Hurd, 2009) and, when asked, people often express their uncertainty about probability distributions (Bruine de Bruin, Fischhoff, Millstein, and Halpern-Felsher, 2000). Imprecise belief measures translate into low explanatory power of economic models for decisions (Drerup, Enke, and von Gaudecker, 2017). Belief dispersion is high even in contexts where private information should not play a major role (e.g. Manski, 2004).

Consequently, there has been a proliferation of theoretical (Ghirardato and Marinacci, 2001; Klibanoff, Marinacci, and Mukerji, 2005; Chateauneuf, Eichberger, and Grant, 2007) and empirical work ((see, e.g., Butler, Guiso, and Jappelli, 2014; Trautmann and van de Kuilen, 2015; Li, Müller, Wakker, and Wang, 2018)) regarding decisions in situations of ambiguity, i.e., those where subjects are uncertain about the correct probability distribution to employ. Overall, we still know much less about ambiguity preferences than about attitudes towards risk or discounting behaviour. Empirical studies have been largely confined to eliciting ambiguity based on Ellsberg (1961), which involves choices about artificial events of unknown distributions (Dimmock, Kouwenberg, Mitchell, and Peijnenburg, 2015; Dimmock, Kouwenberg, and

* We gratefully acknowledge support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC-TR 224 (Project C01) and under Germany's Excellence Strategy – EXC 2126/1– 390838866.

Wakker, 2015; Dimmock, Kouwenberg, Mitchell, and Peijnenburg, 2016; Bianchi and Tallon, 2018; Delavande, Ganguli, and Mengel, 2019).

We add to a recent literature that aims to measure ambiguity attitudes for natural events (Abdellaoui, Baillon, Placido, and Wakker, 2011; Baillon and Bleichrodt, 2015; Baillon, Bleichrodt, Keskin, Haridon, and Li, 2018; Anantanasuwong, Kouwenberg, Mitchell, and Peijnenberg, 2019). Ours is the first study to examine incentivised measures of ambiguity attitudes towards natural events in a representative sample over time and across domains.

To measure ambiguity, we adapt the design of Baillon, Huang, Selim, and Wakker (2018) for use in a representative survey. Using high-powered financial incentives, we elicit four waves of ambiguity attitudes in the context of the stock market over a span of two years; in the fourth wave, we additionally elicit measures from the domain of climate change. Beyond a base payment for survey participation, each individual could earn €20 per wave. Depending on individuals' choices, payment was based on the evolution of a stock market index over the subsequent six month-period or the outcome of a lottery. Expected incentive payments for a rational decision-maker using empirical frequencies for stock returns were €13.5, or an hourly wage of €51.4. On average, we have 92 (minimum: 21, maximum 116) binary decisions at the individual level.

Subjects make several binary decisions between an option whose payoff depends on the development of the stock market and a risky option whose payoff occurs with a known probability. Varying the probabilities of the risky option reveals an individual's *matching probability*; the probability of the lottery at which the subject is indifferent between the two options. We elicit matching probabilities for seven events that depend on the development of the stock market. The distribution of matching probabilities has three salient features. First, the sum of the matching probabilities for an event and its complement is clearly less than one. This indicates that on average, subjects are averse to ambiguity. Second, average matching probabilities are *sub-additive*, the sum of matching probabilities of two mutually exclusive events exceeds the matching probability of their union. This implies individuals are ambiguity-averse for high-probability events and ambiguity-seeking for low-probability events on average (see also Wakker, 2010; Enke and Graeber, 2019). Third, a non-negligible fraction of choice patterns cannot be rationalised by any deterministic theory of choice under uncertainty that we know of. In particular, 57% of subjects at some point assign a higher matching probability to an event that is a strict subset of another.

Based on these observations, we build a model that extracts individual ambiguity attitudes from observed choices whilst accounting for decision errors. Choices depend on three parameters: Ambiguity aversion, which is the average difference between subjective probabilities and matching probabilities. Likelihood insensitivity, which measures how strongly matching probabilities react to underlying changes in subjective probabilities, which can also be interpreted as the perceived level of ambiguity.

Finally, the variance of a random component that affects choices for each event. We structurally estimate these three parameters for each respondent using individual-level choices.

Our first conclusion from this exercise is that ambiguity attitudes are very heterogeneous between respondents, each parameter takes on values within its entire domain. Within respondents, parameters are quite stable; wave to wave correlations average 0.25 for ambiguity aversion and 0.31 for likelihood insensitivity. This is comparable to the stability of risk preferences over similar time spans (Chuang and Schechter, 2015). Within-respondent variation in ambiguity attitudes exhibits no systematic trend over time and bears no meaningful relation to observed characteristics. We interpret this variation as being driven by random fluctuations around a stable mean and by measurement error, which is very prevalent in similar tasks (Gillen, Snowberg, and Yariv, 2018).

Across domains, ambiguity attitudes are more stable than previously thought. The panel dimension of our data allows us to adjust for attenuation due to measurement error by instrumenting parameter estimates with those of previous waves. We find that ambiguity aversion is completely transferable between the domains of finance and climate change but that likelihood insensitivity is not. Our results thus suggest that ambiguity aversion is a domain-invariant preference parameter but that likelihood insensitivity consists of both a transferable and a domain-specific component, which aligns well with the interpretation according to which likelihood insensitivity is the perceived level of ambiguity.

To describe between-respondent heterogeneity in the three dimensions of ambiguity aversion, likelihood insensitivity and the variance of decision errors, we re-estimate the model, pooling data across all waves and assign individuals into groups based on the *k*-means algorithm. Four groups suffice to highlight the most important differences in ambiguity attitudes and their correlates. Almost thirty per cent of the subjects are characterised by a high level of perceived ambiguity and ambiguity aversion. Females, individuals with lower numeracy, higher levels of risk aversion, lower wealth and individuals who perceive positive stock market returns to have occurred less frequently are more likely to belong to this group. Nearly a fifth of participants perceive a similar level of ambiguity but are ambiguity-seeking, not averse. They differ from individuals of the first group in that they are less risk-averse and hold more financial assets. The next group, a third of the population, perceives little ambiguity and is neutral towards it, coming close to expected utility maximising behaviour. High probability and financial numeracy, substantial financial assets and thinking historical returns have often been positive are predictive of belonging to this group. The final group, less than a fifth of subjects, makes more erratic decisions, which prevents reliable measurement of their ambiguity attitudes. Individuals in this group tend to be older, male, have lower rates of numeracy and less knowledge of historical stock returns.

In the next section, we describe the data, the design of our survey instrument and we develop stylised facts that motivate our model. We discuss the identification and estimation of our model in Section 2.3. Section 2.4 presents our estimation results; we conclude in Section 2.5.

2.2 Data, design, and stylised Facts

Our data originate from the LISS panel (Longitudinal Internet Studies for the Social Sciences), an online household panel representative of the Dutch population. Participants answer questionnaires exclusively reserved for research every week and are financially compensated for all questions they answer. Our sample consists of the financial deciders within each household.
In this section, we first present the available background information in the LISS panel, some of which was tailored to our application. Then, we describe our design and highlight some regularities in the choice data it produces.

2.2.1 Background characteristics

In the LISS panel, a variety of information about the households including detailed background characteristics and wealth data is elicited yearly or bi-yearly. Table 2.1 shows the demographics of our sample. The gender split is even. In terms of age, the fraction of 45 to 65-year-olds in our sample is 36 % which is similar to the population-based on aggregate data from Statistics Netherlands (CBS), excluding individuals aged below 20. We have fewer individuals aged 20 to 45 than in the population (25 % compared to 40 %) and more aged 65 to 85 (33 % compared to 18 %). Our sample is also somewhat better educated, with the top two categories of education equalling 13 % and 28 % compared to 11 % and 19 % in the population. These age and education discrepancies with the population are to be expected given that our sample consists of the financial deciders in each household. Income and financial assets are pooled within households. Mean yearly income is close to €28700, mean financial assets are €54800. These are close to the population-wide household numbers in 2018 which were €29500 and €57800 respectively.
During our data collection, as well as in an extra wave in January 2019, we elicited several additional measures to better understand potential drivers of heterogeneity in ambiguity attitudes:

Risk Aversion. One important characteristic that might be related to ambiguity attitudes is risk aversion. We measure households' risk aversion with a preference survey module designed by Falk, Becker, Dohmen, Huffman, and Sunde (2016) which includes a qualitative component, a general risk question, and a quantitative component that is based on elicited certainty equivalents for risky lotteries. We combine

Table 2.1. Summary statistics

	Observations	Mean	Std. dev.	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$
Female	2235	0.5				
Education: High school or less		0.35				
Education: Junior college		0.24				
Education: College		0.28				
Education: University		0.13				
Age: $\in (20, 45]$	2230	0.25				
Age: $\in (45, 65]$		0.37				
Age: $\in (65, 80]$		0.33				
Age: > 80		0.05				
Income (thousands)	1806	2.39	1.49	1.5	2.17	3
Financial assets (thousands)	1838	54.77	157.38	2.42	15	46.69

Notes: Sample restrictions: Individuals with at least two waves of regular choices in the ambiguity tasks. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85 % of subjects. Income and assets are pooled within households, data from 2018.

the quantitative and qualitative components as suggested in Falk, Becker, Dohmen, Huffman, and Sunde (2016).

Numeracy. The ability to reason quantitatively is particularly important when making decisions under uncertainty. We measure three components of numeracy. First, a basic numeracy component based on the English Longitudinal Study of Ageing (Steptoe, Breeze, Banks, and Nazroo, 2013). Second, a financial numeracy component that involves interest rates and inflation for which we used a subset of the questions by van Rooij, Lusardi, and Alessie (2011). Third, a probability numeracy component that tests both basic understanding of probabilities and more advanced concepts such as independence and additivity. We use the questions proposed by Hudomiet, Hurd, and Rohwedder (2018) and additionally add two questions that could be particularly informative about the types of errors that can occur when individuals make decisions in our design. We aggregate the three components into a numeracy index, giving equal weight to each component.

Our measures of numeracy and risk aversion are related to socio-demographics characteristics as in the previous literature (e.g., Dohmen, Falk, Huffman, Sunde, Schupp, et al., 2011; van Rooij, Lusardi, and Alessie, 2011; Hudomiet, Hurd, and Rohwedder, 2018): Older, less educated, and female subjects tend to have lower numeracy skills and are more risk-averse (Table 2.B.2).

Judged historical frequencies of past AEX returns. We also asked individuals to judge how frequently the AEX events used in our designs occurred over the last 20 years. Although there is substantial individual heterogeneity, the last column of Table 2.2 shows that the average judged frequencies are not too far from the empirical frequencies. Subjects underestimate the frequency of positive returns on average but think returns greater than 10 % occurred more often than they did.

Optimism. Optimism is a potential determinant of ambiguity attitudes. We elicited optimism and pessimism measures based on the revised life orientation test (Scheier, Carver, and Bridges, 1994), combining them into an overall measure of optimism.

Knowledge of and concern about climate change. To help analyse ambiguity attitudes towards climate change, we asked subjects to report (i) their understanding of the causes and implications of climate change on a five-point scale and (ii) whether climate change is a threat to them and their family on a six-point scale.

2.2.2 Measuring Ambiguity Attitudes

Our goal is to investigate the distribution and stability of ambiguity attitudes in a representative population. In our main application, we choose the stock market as the source of uncertainty, since decisions under ambiguity are very prevalent in this domain. Furthermore, the subjects are unable to influence the outcome in this context which allows for the incentivisation of their choices. As a benchmark for the stock market, we employ the Amsterdam Exchange Index (AEX), the most important stock index in the Netherlands. Individuals make several binary decisions between an option whose payoff depends on the development of the AEX over the next six months and an option whose payoff occurs with a known probability.

When measuring ambiguity attitudes about natural events, the challenge is to control for any subjective beliefs individuals may hold about them. Suppose we observe individuals refrain from tying their payoff to an increase of the AEX index. This could be either because they perceive AEX returns as ambiguous and are averse to such ambiguity, or because they consider positive AEX returns to be unlikely. To disentangle the two explanations based only on observed choices, we use the design of Baillon, Huang, Selim, and Wakker (2018) in which the role of subjective beliefs is neutralised by having individuals make decisions about events and also the complement of events.

One example of a binary choice situation that forms the core of our design is visualized in Figure 2.1. Option 1 pays twenty Euros if the performance of a hypothetical €1000 investment in the AEX over the next six months is within a certain range. In this example, twenty Euros will be paid if the investment is worth more than €1100 in six months, i.e. an increase of more than 10 %. Option 2 is a lottery and pays twenty Euros with probability 50 %, visualised by a pie chart.

Multiple choices between such options provide information about the *matching probability* an individual assigns to the AEX event, which is defined as follows:

Definition 2.1 (Matching probability)**.** The matching probability $m(E)$ of an event E is the probability p that makes a decision-maker indifferent between a pay-out of X if event E occurs and a bet on a lottery that pays X with probability p.

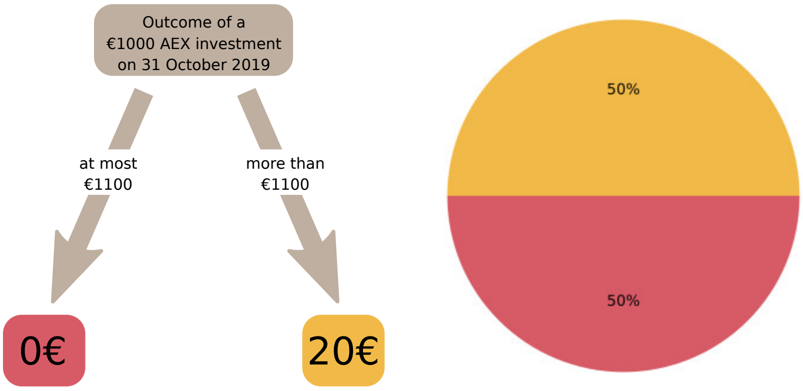


Figure 2.1. Exemplary binary choice situation: ambiguous option and risky option

Notes: Labels are translated from Dutch to English.

A chained design of 3–4 binary choices is used to identify the matching probability of an event. Compared to a choice list format, we expect this procedure to reduce complexity for the subjects as they can focus on one question at a time. After every choice, the probability of Option 2 changes depending on the previous choice, pinning down the matching probability to within 0.1. The complete decision tree is shown in Figure 2.A.1.

Following the logic of the design by Baillon, Huang, Selim, and Wakker (2018), we partition the space of possible values the AEX investment can take into three events: $E_1 : Y_{t+6} \in (1100, \infty]$, $E_2 : Y_{t+6} \in (0, 950)$, and $E_3 : Y_{t+6} \in [950, 1100]$. We chose this partition to balance historical 6-month returns of the AEX, for which the respective frequencies were 0.24, 0.28 and 0.48. We elicit matching probabilities for each of these events as well as their complements but initiate respondents by having them first consider the more intuitive event $E_0 : Y_{t+6} \in (1000, \infty]$, i.e. that the value of the investment will increase. The resulting seven events for which we elicit matching probabilities are depicted in Figure 2.2.

Because eliciting attitudes about ambiguous events comes with a substantial cognitive burden for participants, we try to make the design as easy to comprehend as possible. We included a tutorial in the design that introduces participants to the choices and their potential consequences.

To analyse stability over time, we repeat the elicitation procedure just described four times. The design was semi-annually rolled out alongside the regular core questionnaires of the LISS panel. We have collected data from waves in May 2018, November 2018, May 2019, and November 2019. Originally, 2773 financial deciders were invited to participate, of which 2146, 2170, 2000, and 1957 completed the questionnaire in the respective waves. One of the binary choices in every wave is played

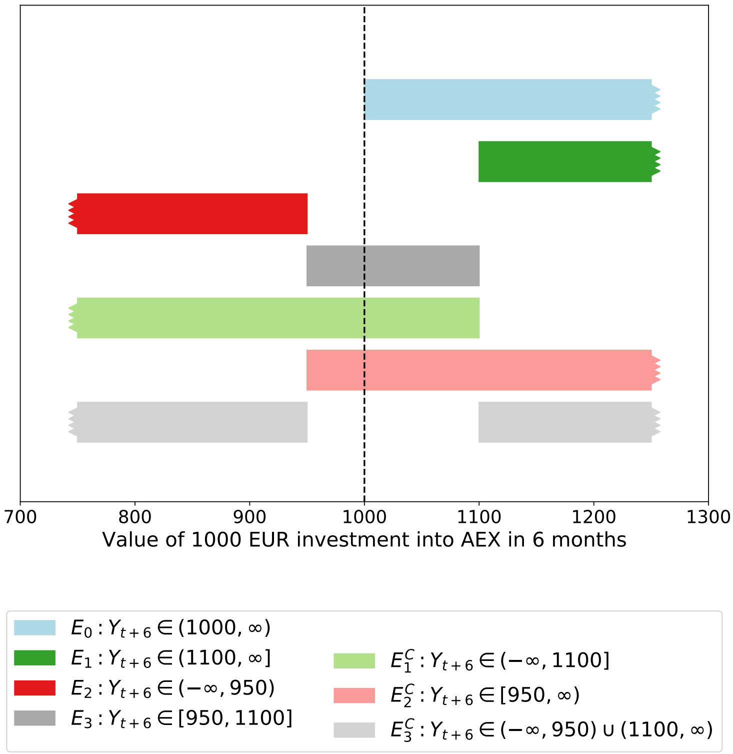


Figure 2.2. Events of AEX performance used in the experiment

out half a year later, at the start of the next questionnaire, with a possible pay-out of twenty Euros depending on the development of the AEX and chance.[1]

2.2.3 Matching Probabilities and Errors

Next, we analyse the distribution of matching probabilities and develop several insides that we use later to build up the empirical model.

Some individuals pick the same option throughout an entire wave, i.e. 28 times in a row. This behaviour could be interpreted as an extreme form of ambiguity aversion or ambiguity seeking but an alternative explanation is that some individuals do

[1] Because the choice at each node determines the options at the subsequent node, the design would not be incentive compatible if we selected one of the answered questions for pay-out ex-post. To circumvent this problem, the question that is paid out is selected out of all 91 possible choice situations before the specific subject made any decisions. If the subject did not encounter the selected question during the questionnaire because it was in a different branch of the decision tree, the question is additionally asked at the end of the questionnaire. This mechanism is inspired by Johnson, Baillon, Bleichrodt, Li, Dolder, et al. (2015) and has been implemented in a similar fashion by Bardsley (2000). The fact that the choice that is paid out is pre-selected also prevents the subjects from hedging against the encountered ambiguity (Baillon, Halevy, and Li, 2014)

not seriously contemplate the choices. As Figure 2.A.2 shows, many of these subjects go through the questionnaire much faster than the rest which points to the latter explanation. We drop subjects if two conditions are met. First, their answers exhibit such patterns and second, their average response time on the first decision for each event is below the 15th percentile of all subjects. We exclude such individuals from the analysis on a wave by wave basis which decreases the sample size by 2.5 %.

As the mean of matching probabilities within events is fairly stable across waves (see Table 2.B.1), Table 2.2 depicts summary statistics of the elicited matching probabilities pooled across all waves. The last two columns show the empirical frequencies with which the events occurred and the mean judged historical frequencies reported by the subjects.

Table 2.2. Matching probabilities, empirical frequencies and judged historical frequencies

	Mean	Std. dev.	$q_{0.1}$	$q_{0.5}$	$q_{0.9}$	Empirical Frequency	Judged Hist. Frequency
$E_0 : r > 0\%$	0.51	0.28	0.08	0.45	0.92	0.63	0.52
$E_1 : r > 10\%$	0.35	0.25	0.03	0.35	0.75	0.24	0.31
$E_1^C : r \leq 10\%$	0.53	0.29	0.15	0.45	0.97	0.76	
$E_2 : r < -5\%$	0.37	0.27	0.03	0.35	0.75	0.28	0.22
$E_2^C : r \geq -5\%$	0.54	0.30	0.08	0.55	0.97	0.72	
$E_3 : -5\% \leq r \leq 10\%$	0.57	0.29	0.15	0.55	0.97	0.48	0.47
$E_3^C : (r < -5\%) \cup (r > 10\%)$	0.41	0.28	0.03	0.35	0.85	0.52	

Notes: Events were asked about in this order: $E_0 - E_1 - E_2 - E_3 - E_1^C - E_2^C - E_3^C$. Summary statistics for the matching probabilities of the seven events are shown. Matching probabilities are set to the midpoint of the interval identified by the design. Data is pooled across all waves. The last two columns show the empirical frequencies (starting from 1992, own calculation) and the mean judged historical frequencies (reported by the subjects). Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects.

Three observations can be made. First, the sum of the average matching probabilities of an event and its complement event, e.g. E_1 and E_1^C is less than 1 for all three events E_1, E_2 and E_3. This is an indication that matching probabilities do not equal subjective probabilities, and that individuals experience ambiguity aversion on average. Second, mean matching probabilities are on average *sub-additive,* in the sense that the sum of the matching probabilities of E_1 and E_2 is well above the average matching probability of their union, E_3^C. The same relation is found for the other combinations of two single events. Third, the average matching probability for E_3 is slightly larger than E_1^C or E_2^C. This is surprising because E_3 is a subset of the other two events, and subsets cannot be considered more attractive bets than supersets under any reasonable theory.

If this *set-monotonicity* requirement is violated, it is an indication of a decision error. There are eight pairs of events at which such an error can occur. In total, 57 %

of individuals violate set-monotonicity at least once in a given wave. The rate of set-monotonicity violations for a given superset-subset pair depends on the difference of judged historical frequencies of the two events – both in the aggregate (Table 2.B.3) and on the individual level (Table 2.B.4). This is an indication that erroneous answers are not driven by random behaviour alone, but also depend on the events involved. This informs how we specify the error component in our model.

To analyse stability across domains, in the 4th wave we additionally elicit ambiguity attitudes in another domain: climate change. We use the same setup as before, replacing events relating to the AEX with events relating to mean temperature changes during the winter 2019/2020 compared to the previous five winters. The possible temperature changes are partitioned into three events, using cut-offs at +1°C and -0.5°C. We elicit matching probabilities for the three single events, the three complementary events, and the additional event that the temperature change is at least +0°C. Table 2.B.5 shows summary statistics of the matching probabilities.

2.3 Empirical strategy

Based on the observations in the last section, we now introduce the empirical model we use to estimate ambiguity attitudes.

2.3.1 Defining and interpreting ambiguity attitudes

We build upon the bi-separable utility framework of Ghirardato and Marinacci (2001). In that framework, a prospect that pays out X if event E occurs and otherwise nothing is evaluated as $W(E) \cdot V(x)$ where $V(\cdot)$ can be any utility function and $W(\cdot)$ a *decision weight*. W satisfies the following conditions $W(\emptyset) = 0$, $W(\Omega) = 1$, and $B \subseteq A \implies W(B) \leq W(A)$. We assume the decision weight depends on the subjective probability agents assign to the event, where the relation between the two is governed by a *source function* w_S such that $W(E) = w_S(\Pr(E))$ (Abdellaoui, Baillon, Placido, and Wakker, 2011).[2] The subscript indicates that the function depends on the source of uncertainty, which is the mechanism that generates it. In this paper, we examine uncertainty about the future development of the AEX and uncertainty about temperature changes.

A subject is ambiguity-averse for an event E if $W(E) < \Pr(E)$, ambiguity-neutral if $W(E) = \Pr(E)$, and otherwise ambiguity-seeking. There is empirical evidence that the degree of ambiguity aversion about an event varies with the subjective probability the decision-maker assigns to it. When individuals stand to gain if an uncertain event occurs, the most common pattern is ambiguity seeking for events individuals

[2] Individuals can be thought of as having subjective probabilities in mind or as making choices that can be rationalized with them.

regard as long shots and ambiguity aversion for medium or high probability events (Trautmann and van de Kuilen, 2015).

To capture both average ambiguity aversion and its dependence on the subjective probability, we specify $w_S(\Pr(E))$ as the *neoadditive* function introduced by Chateauneuf, Eichberger, and Grant (2007) which has been shown to fit choices very well in settings where both decisions and $\Pr(E)$ are observed (Li, Müller, Wakker, and Wang, 2018).

$$\begin{aligned} W(E) &= \tau_0 + \tau_1 \cdot \Pr(E), \text{ for } \Pr(E) \in (0,1) \\ W(\varnothing) &= 0, W(\Omega) = 1 \\ 0 &\leq \tau_1 \leq 1, 0 \leq \tau_0 \leq 1 - \tau_1 \end{aligned}$$

The conditions on the parameters ensure that $W(E)$ equals 0 and 1 only for events agents regard as impossible or certain, unless $\tau_1 = 0$ and subjective probabilities play no role at all. They also rule out that individuals assign a greater weight to events they regard as less probable[3].

In terms of τ_0 and τ_1 we can define two ambiguity parameters:

$$\text{Ambiguity aversion} \quad \alpha = \frac{1 - 2\tau_0 - \tau_1}{2} = E[\Pr(E) - W(E)] \tag{2.1}$$

$$\text{Likelihood insensitivity} \quad \ell = 1 - \tau_2 = 1 - \frac{Cov(W(E), \Pr(E))}{Var(\Pr(E))} \tag{2.2}$$

Ambiguity aversion is the average amount by which subjective probabilities exceed decision weights, where we average across all subjective probabilities in the unit interval with equal weight. For the neoadditive function, this is equivalent to $Pr(E) - W(E)$ at $\Pr(E) = 0.5$. Likelihood insensitivity captures the extent to which individuals' decision weights change if the underlying subjective probabilities change. This is 1 minus the slope of the source function, $1 - \tau_1$. Figure 2.3 illustrates the concepts for $\alpha = 0.1$ and $\ell = 0.6$. Lower τ_1 and therefore higher ℓ corresponds to a flatter function, i.e. event weights are less responsive to subjective probabilities. An increase of α, on the other hand, corresponds to a downwards shift of $W(E)$ for all subjective probabilities. The range of possible values for α is determined by the level of ℓ. Only for $\ell = 1$, the maximum level of ambiguity aversion ($W(E) = 0$ for all $\Pr(E) \in (0,1)$) or the maximum level of ambiguity seeking ($W(E) = 1$ for all $\Pr(E) \in (0,1)$) can be detected. On the other hand, $\ell = 0$ ensures $W(E) = \Pr(E)$, which is the case of expected utility maximisation.

[3] In the previous section, we documented that there are set-monotonicity errors for a sizable fraction of individuals, which is an example of giving greater weight to an event that must be less probable. This is one of the reasons we augment the deterministic neoadditive model with a random error component when we estimate it.

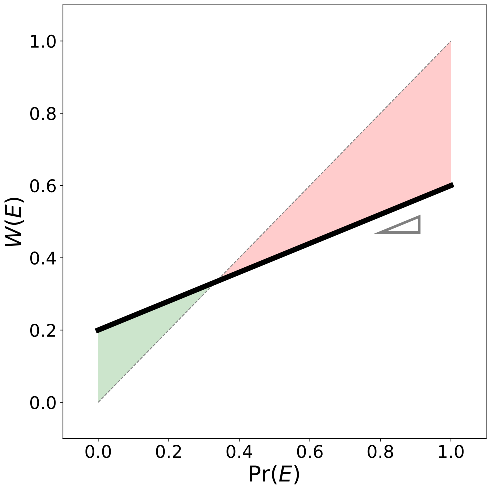


Figure 2.3. Ambiguity aversion and likelihood insensitivity with a neoadditive source function

Notes: The figure plots the neoadditive source function $W(E) = \frac{\ell}{2} - \alpha + (1 - \ell) \cdot \Pr(E)$ for $\alpha = 0.1$ and $\ell = 0.6$. Ambiguity aversion α is the red area between $\Pr(E) - W(E)$ where the difference is positive less the green area where the difference is negative. It also equals the distance $Pr(E) - W(E)$ at $\Pr(E) = 0.5$. Likelihood insensitivity is 1 minus the slope of the source function (black line) which is indicated by a grey triangle.

In addition to its interpretation as part of a plain decision weight, ℓ can also be regarded as the perceived level of ambiguity due to the role it plays in multiple prior models (Chateauneuf, Eichberger, and Grant, 2007; Baillon, Bleichrodt, Keskin, Haridon, and Li, 2018). In such a model, individuals evaluate a bet on E with a weighted average of expected utilities calculated with the least and most optimistic belief in an interval of priors. ℓ is the width of the interval and $0.5 + \frac{\alpha}{\ell}$ the weight of the pessimistic expected utility term[4]. This interpretation requires that $\ell \geq 0$ because otherwise the width of the interval would exceed 1, and that $-\frac{\ell}{2} \leq \alpha \leq \frac{\ell}{2}$ for the utility term weights to be in [0, 1].

These conditions are enforced in our main specification and correspond to the conditions on τ_0 and τ_1 stated earlier. While the violation of set-monotonicity ($\ell > 1$) is incompatible with any reasonable model of decision making, the plain decision weight interpretation allows for behaviour such as $\ell \leq 0$ which we might interpret as being hypersensitive to subjective probabilities. In appendix 2.C, we estimate our model keeping only the restriction $\ell \leq 1$, which means that ℓ cannot necessarily be interpreted as the perceived level of ambiguity although the decision weight interpretation remains intact. We find that the estimated ambiguity attitudes of only 12% of individuals fall outside the restrictions of our main specification and that our key results are unaffected.

[4] Except for $\ell = 0$, the expected utility case, when the weights are 0.5

2.3.2 Estimating ambiguity attitudes

Since matching probabilities find the indifference point $W(E) \cdot V(€20) = p \cdot V(€20)$, they identify the decision weight individuals assign to AEX events when making decisions relating to them: $W(E) = m(E) = p$.[5] The decision weights are identified independently of the functional form of the utility function and, in particular, independently of risk aversion.

It is easy to see that the neoadditive model, and hence α and ℓ, are identified in terms of the matching probabilities for the events in our design: The difference between $W(E_1) + W(E_2) + W(E_3) = 3\tau_0 + \tau_1$ and $W(E_j) + W(E_j^C) = 2\tau_0 + \tau_1$ identifies τ_0, and then τ_1 is also identified. The subjective probabilities drop whatever they are because the events in the design contain their complements as well.

To capture erratic behaviour as well as systematic behaviour that is not captured well by the deterministic neoadditive model, we augment it with an additive error ϵ_E which we assume is normally distributed with mean zero and a standard deviation of σ_ϵ independently across events. An additive error for events is motivated by the finding documented in Section 2.2.3 that set-monotonicity violations are related to differences in judged historical frequencies of the respective events: Errors are more likely if individuals believe that a pair of events forming a superset and subset have occurred similarly often in the past. Errors that are not specific to events, such as trembling hand errors, cannot generate this pattern.

[5] We implicitly assume that there is no probability weighting for known probabilities and, hence, $w_{risk}(p) = p$. If this not the case, our results are still informative about ambiguity attitudes in that they measure the difference in weights under uncertainty and risk.

We estimate the following model

$$\begin{aligned} W(E) &= \tau_0 + \tau_1 \cdot \Pr(E) \\ \epsilon_E &\sim \mathcal{N}\left(0, \sigma_\epsilon^2\right) \\ \Pr(\text{Observed choices for E}) &= \Pr(W(E) + \epsilon_E \in \text{Interval}_E) \end{aligned}$$

by choosing the parameters $\theta := [\tau_0, \tau_1, \Pr(E_0), \Pr(E_1), \Pr(E_2), \sigma_\epsilon]$ to maximise the likelihood

$$\begin{aligned} \mathscr{L}(\theta) &= \prod_{E \in \mathfrak{E}} \Pr(W(E) + \epsilon_E \in \text{Interval}_E; \theta) \\ &\text{s.t.} \quad 0 \leq \tau_1 \leq 1, \; 0 \leq \tau_0 \leq 1 - \tau_1, \\ &\Pr(E_1) \leq Pr(E_0), \; Pr(E_1) + Pr(E_2) \leq 1, \; \Pr(E) \in [0,1] \end{aligned}$$

for the events E in $\mathscr{E} = \{E_0, E_1, E_2, E_3, E_1^C, E_2^C, E_3^C\}$. Pr(Observed choices for E) is the probability of the sequence of observed choices regarding event E, all of which lead to one of the terminal intervals depicted in Figure 2.A.1.

Baillon, Bleichrodt, Li, and Wakker (2019) propose indices that estimate α and ℓ directly with moments of matching probabilities. Our approach is more difficult to implement because it requires solving constrained optimisation problems for each individual, but it gives us several advantages. First, it ensures that estimated ambiguity parameters obey the theoretical parameter restrictions that rule out irrational behaviour and allow ℓ to be interpreted as the perceived level of ambiguity. Figure 2.D.3 shows the distribution of estimated parameters when the estimation is based on the indices of Baillon, Bleichrodt, Li, and Wakker (2019). For 25% of subjects, the estimates of ℓ are above 1 which implies they give lower weights to events with higher subjective probabilities. Rather than excluding these individuals or disregarding that such parameter values are not meaningful, we find the best fitting parameters subject to their values being interpretable.

Second, we obtain an extra parameter σ_ϵ. This error parameter informs us about the fit of the model for each subject's choices and therefore the reliability with which α and ℓ are estimated. Individuals that frequently violate set-monotonicity, for instance, will have a high value of σ_ϵ. Third, our approach allows us to use choices for the seventh event E_0 when estimating ambiguity parameters which improves efficiency. These choices could only be included in the indices if choices for the complement event were available as well.

Finally, note that estimating the neoadditive model entails little loss of generality compared to the indices from a theoretical perspective. The indices are invariant to the choice of events in the design only if the neoadditive model is true and ℓ is estimated well if the neoadditive model is a good approximation of the source function (Baillon, Bleichrodt, Li, and Wakker, 2019, Theorem 14 and Proposition 21). Using σ_ϵ, we can quantify the quality of the approximation for each individual. Appendix 2.D repeats our empirical analysis with the indices and comes to broadly similar conclusions, but estimates of ℓ are substantially less stable over time and across domains compared to estimates from our model.

2.4 Results

We now present our results about the estimated ambiguity parameters. First, we examine stability over time, as well as stability across domains. In the last part of this section, we assess the heterogeneity of our three parameters using a discrete classification approach.

2.4.1 Parameter stability over time

To examine the stability of estimated ambiguity attitudes over time, we make use of the panel structure of our data and estimate our model separately for each individual and survey wave. Figure 2.4 shows boxplots of the distribution of parameter estimates for each wave. The shapes of the distributions are quite stable wave to wave, particularly those of the ambiguity parameters α and ℓ. The distribution of σ_ε, however, noticeably shifts to the left following the first wave and seems to stabilise thereafter. The reduction of the error parameter likely reflects both a small change in the experimental instructions that made the description more intuitive and a greater familiarity of the respondents with our design.

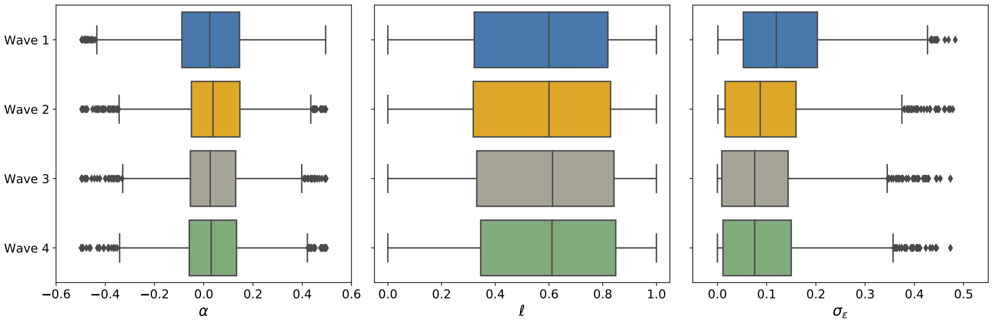


Figure 2.4. Distributions of estimated parameters, wave by wave

Notes: Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects.

To check whether there might be systematic heterogeneity in changes over time that cancels out in the aggregate analysis, we regress changes in estimated parameters across consecutive survey waves on many observables. The results in Table 2.B.6

show that parameter changes are only very weakly related to observable characteristics, with R^2 below 1% for the ambiguity parameters. There is little evidence in our data that ambiguity parameters are systematically changing over the two years.

Figure 2.4 also shows that there is substantial variation in all estimated parameters. The ambiguity parameters are spread over the full range of their support. To investigate individual-level parameter stability, we compute correlations between parameter estimates for all pairs of survey waves. Table 2.3 shows the results. On average, correlations are 0.25 for ambiguity aversion and 0.31 for likelihood insensitivity though they tend to be higher for consecutive survey waves, which are six months apart, and between survey waves not involving the first wave which was the first exposure of individuals to our design. To interpret the magnitude of these correlations, a comparison with results on risk aversion is instructive. Chuang and Schechter (2015) review the literature on the stability of risk aversion parameters over longer horizons comparable to ours, finding correlations between 0.13 and 0.55 for studies with at least 100 observations. Our results indicate that ambiguity attitudes are of comparable stability to risk attitudes.

Table 2.3. Across wave correlations of estimated parameters

	$\hat{\rho}_{1,2}$	$\hat{\rho}_{1,3}$	$\hat{\rho}_{1,4}$	$\hat{\rho}_{2,3}$	$\hat{\rho}_{2,4}$	$\hat{\rho}_{3,4}$	Average $\hat{\rho}$
α	0.25	0.22	0.20	0.26	0.21	0.33	0.25
ℓ	0.24	0.22	0.28	0.35	0.36	0.42	0.31
σ_ε	0.16	0.20	0.21	0.32	0.32	0.36	0.26

Notes: Table shows Pearson correlations of parameter estimates between the survey waves indicated by the subscripts. Parameter estimates are obtained from the model described in Section 2.3.2 separately for each survey wave and individual. Sample restrictions: Observations with regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects.

The moderate magnitude of the correlations means that there is substantial variation in estimated parameters within individuals. As Schildberg-Hörisch (2018) points out regarding risk preferences, this variation likely reflects both measurement error and temporary fluctuations of the underlying parameter around each individual's mean level of the parameter. To address measurement error, we adopt two approaches: When examining stability across domains, we instrument estimated parameters with estimated parameters of other waves. For Section 2.4.3, in which we analyse between-subject heterogeneity, we re-estimate our model, pooling individual choices across survey waves.

2.4.2 Parameter stability across domains

A key question arising for any parameter characterising individual attitudes is how domain-specific it is. Do attitudes towards uncertainty about how the AEX will evolve extend to other, non-financial domains? To address this question, we elicited

α and ℓ not only for events relating to the AEX but also to events relating to how the average temperature in the winter of 2019 compares to the previous five years. Figure 2.5 compares the respective distributions of parameters in wave 4. For α and σ_ϵ, the distributions are very similar, but there is notably greater likelihood insensitivity regarding temperature changes. In the following, we examine stability at the individual level.

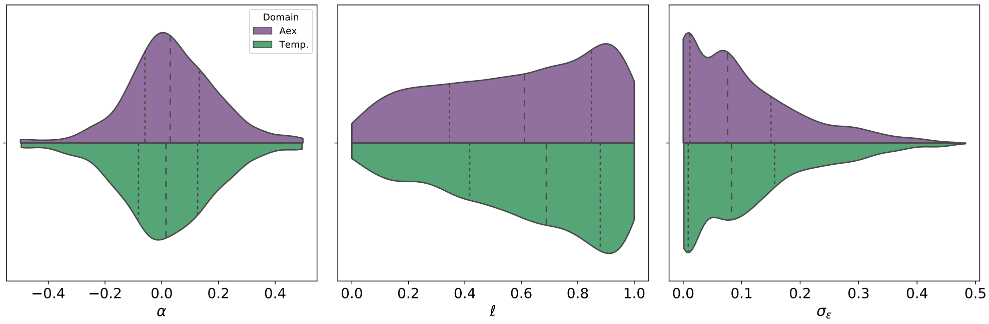


Figure 2.5. Distributions of estimated parameters, financial v climate domains

Notes: Estimates from for both domains use data from wave 4. The dashed line shows the median, the dotted lines bottom and top quartiles. Sample restrictions: Observations with regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects.

Table 2.4 shows regressions for each parameter in the climate domain on parameters from the financial domain elicited in the same wave. The first column of each parameter shows OLS regression with slope coefficients of 0.70, 0.36, and 0.50 for α, ℓ, and σ_ϵ respectively. This suggests a sizable amount of stability across domains, but a much higher stability for ambiguity aversion compared to likelihood insensitivity. The second columns add several controls. For brevity, the coefficients of control variables are shown in the appendix in Table 2.B.8. Our results are unchanged when we control for demographic variables, numeracy, risk aversion, and the extent to which individuals think they understand climate change and deem it a threat. Stability across domains is not driven by these common correlates.

However, the OLS regressions are distorted by estimation error in potentially two ways. First, if estimates of ambiguity attitudes are subject to classical measurement error, the slope coefficients are attenuated to zero and understate the degree to which the parameters are stable across domains. Second, there could be a positive correlation between the estimation errors for estimates across domains, because the parameters were elicited one after another in the 4th wave. This would cause the coefficients to overstate the dependence across domains. To address this, we run two-stage least squares regressions in the third columns for each parameter, instrumenting the AEX related parameters of the 4th wave with those of the previous waves. If estimation errors are uncorrelated across waves, this eliminates both biases.

Table 2.4. Dependence of parameters relating to temperature uncertainty on parameters relating to uncertainty about the AEX

Parameter		α			ℓ			σ_ε	
Model	OLS	OLS	2SLS	OLS	OLS	2SLS	OLS	OLS	2SLS
Intercept	-0.01**	0.05*	0.00	0.42***	0.44***	0.20***	0.06***	0.01	-0.03
	(0.00)	(0.03)	(0.03)	(0.02)	(0.06)	(0.07)	(0.00)	(0.02)	(0.02)
AEX param	0.70***	0.70***	1.00***	0.36***	0.34***	0.61***	0.50***	0.48***	1.06***
	(0.03)	(0.03)	(0.09)	(0.03)	(0.03)	(0.06)	(0.03)	(0.03)	(0.11)
Underst. c.c.		-0.01**	-0.01**		-0.02**	-0.02**		0.01**	0.01***
		(0.00)	(0.00)		(0.01)	(0.01)		(0.00)	(0.00)
Threat. by c.c.		0.00	0.00		-0.00	0.01		0.00	-0.00
		(0.00)	(0.00)		(0.01)	(0.01)		(0.00)	(0.00)
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
N	1297	1297	1186	1297	1297	1186	1297	1297	1186
R^2	0.402	0.416	-	0.146	0.170	-	0.216	0.236	-
1st st. F	-	-	79.8	-	-	308.4	-	-	134.9

Notes: Outcomes are estimated parameters in the temperature domain in the 4th wave, regressors are estimated parameters in the AEX domain in the 4th wave. Two-stage least squares models use estimated parameters from the previous three waves as instruments. Controls are age, gender, education, income and assets dummies, risk aversion, basic, financial and probability numeracy and indicators of self-assessed understanding and perceived threat of climate change with a 5 and 6 point scale respectively (see Table 2.B.8). Robust standard errors in parentheses. Sample restrictions: Observations with regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects.

The regressions that adjust for estimation error strikingly show that ambiguity aversion and the magnitude of errors is completely stable across the two domains with point estimates close to 1. This supports the interpretation of ambiguity aversion as stable preference that fully extends across domains. Anantanasuwong, Kouwenberg, Mitchell, and Peijnenberg (2019) elicit ambiguity attitudes for events from different financial domains: Individual stocks, local and foreign stock indices and crypto funds. They find that ambiguity aversion parameters are very related across these domains with R^2 between 0.4 and 0.54. This is in line with what we find in the OLS regression. A coefficient close to 1 in the 2SLS regression that adjusts for estimation error is likewise in line with the conclusion of Anantanasuwong, Kouwenberg, Mitchell, and Peijnenberg (2019) who conjecture based on a factor analysis that there is only one underlying ambiguity aversion. Our results indicate that the stability of ambiguity aversion holds not just within financial contexts, but more generally.

We further find that ℓ also has a substantial transferable component, but the slope coefficient of 0.60 is well below 1. Based on the multiple prior interpretation of ℓ as the perceived level of ambiguity, this is expected as perceptions are more likely to differ across domains than preferences. This interpretation is strengthened by the fact that self-reported knowledge of climate change has substantial predictive power for the perceived level of ambiguity in the climate domain, conditional on the perceived level of ambiguity in the financial domain. Anantanasuwong, Kouwenberg, Mitchell, and Peijnenberg (2019) find a very weak dependence across domains with R^2 ranging from 0.005 to 0.032 which would imply that ℓ is almost completely context-specific. Our analysis shows that a substantial component of likelihood insensitivity is stable across domains. One potential reason our results on the perceived level of ambiguity are at variance with Anantanasuwong, Kouwenberg, Mitchell, and Peijnenberg (2019) is measurement error. Table 2.4 demonstrates that our model-based estimates are subject to sizable measurement error and there is evidence it affects ambiguity attitudes estimated with indices, instead of our model, even more. In Table 2.D.2 we replicate Table 2.4 with the index-based estimates that Anantanasuwong, Kouwenberg, Mitchell, and Peijnenberg (2019) use, and get a comparably small R^2 of 0.028 for ℓ. The 2SLS-measurement-error-adjusted regression slope is, however, in the range of what we find with our model. In line with this explanation, index-based estimates of ℓ are substantially less stable over time (Table 2.D.1).

Our findings suggest that there can be room for external stimuli, such as providing individuals with more information about a source of uncertainty, to change ℓ while this might not be possible for α. This aligns well with the findings by Baillon, Bleichrodt, Keskin, Haridon, and Li (2018) who conduct such an information experiment.

As with stability over time, the comparison with risk aversion is instructive. Dohmen, Falk, Huffman, Sunde, Schupp, and Wagner (2011) examine self-reported assessments of risk aversion in several domains like financial matters, sports, or health and report correlations that correspond to R^2 between 0.16 to 0.36 which is comparable to what we find in the OLS columns of Table 2.4. Dohmen, Falk, Huffman, Sunde, Schupp, and Wagner (2011) reason that differences in risk attitudes across domains might be more likely to reflects different risk perceptions, rather than differences in actual preferences. This is in line with what we find for ambiguity: A very stable ambiguity aversion component, but that the perception of ambiguity varies across contexts to a certain degree.

2.4.3 Describing heterogeneity in attitudes and error propensities

In this section, we examine heterogeneity in ambiguity attitudes and error propensities and their relation to other individual characteristics. To improve precision, we re-estimate our model, holding ℓ, α, and σ_ϵ fixed but allowing the subjective probabilities to change between waves.

It is crucial to consider the joint distribution of parameters rather than each parameter in isolation for two reasons: First, the error parameter is informative about how reliably the other parameters are estimated, both in terms of statistical precision and fit of the neoadditive model. Second, the magnitude of ambiguity aversion or seeking that can be detected by our design depends on the perceived level of ambiguity. When $\ell = 0$ it must be the case that $W(E) = \Pr(E)$, so there is no scope for ambiguity aversion or seeking.

With this in mind, we classify individuals into one of a discrete set of groups using all three estimated parameters and consider the most striking features of each resulting group. We use the k-means algorithm to do this. For a given number of groups, it assigns individual observations $x_i := [\alpha_i, \ell_i, \sigma_{\epsilon,i}]$ to groups g such that $\sum_i ||x_i - c_{g(i)}||^2$ is minimised for the group means $c_g = \frac{1}{N_g} \sum_{i \in g} x_i$. We scale x_i to mean 0 and standard deviation 1 in the cross section to ensure every component is given equal weight in the optimisation.

We summarise the results of this exercise for $K = 4$ groups, which is the minimum necessary for there to be meaningful group-level differences along the three parameters. In Section 2.E, we double the number of groups and show that the qualitative insights from the $K = 4$ analysis remain intact. We describe the groups in two figures and two tables, with groups sorted by their average ℓ from high to low: Figure 2.6 shows the distribution of ambiguity profiles in $(\alpha,\ \ell)$ with the large diamonds indicating group means and the small dots indicating individual profiles. Figure 2.7 shows the source function (how decision weights depend on subjective probabilities) for the average ambiguity profile of each group, as well as the average magnitude of the error component. Table 2.5 lists means and medians of observable characteristics per group and Table 2.6 displays marginal effects of a multinomial logit regression predicting group membership based on the same characteristics.

Group 1: Substantial likelihood insensitivity, on average ambiguity-averse. Almost thirty per cent of individuals in our sample show substantial likelihood insensitivity with ℓ ranging from 0.4 to 1, and are averse to it, with α ranging from 0 to 0.5. Their choices are quite consistent with the neoadditive model, leading to a comparably small error magnitude of 0.14. The blue line in Figure 2.7 crosses the 45-degree line just before the subjective probability reaches 0.3 and rises only up to a matching probability of about 0.5. This means on average, individuals in this group are quite ambiguity-averse; they prefer bets on lotteries over bets on AEX events even if they regard them as substantially more likely. In Table 2.5 we see that individuals of Group 1 are likely to be somewhat younger than those of other groups, and more likely to be female. They tend to be more risk-averse and hold substantially less financial assets. Besides, group 1 individuals are on average less optimistic than those of groups 2 and 4, both in terms of a personality measure and in terms of how often they think the AEX had a positive return over the last 20 years. Except age and optimism, the characteristics mentioned are also predictive of membership in group 1 in a multinomial logistic regression (Table 2.6).

Group 2: Substantial likelihood insensitivity, on average ambiguity-seeking. A smaller group, a fifth of individuals, is associated with a similar ℓ as group 1 and

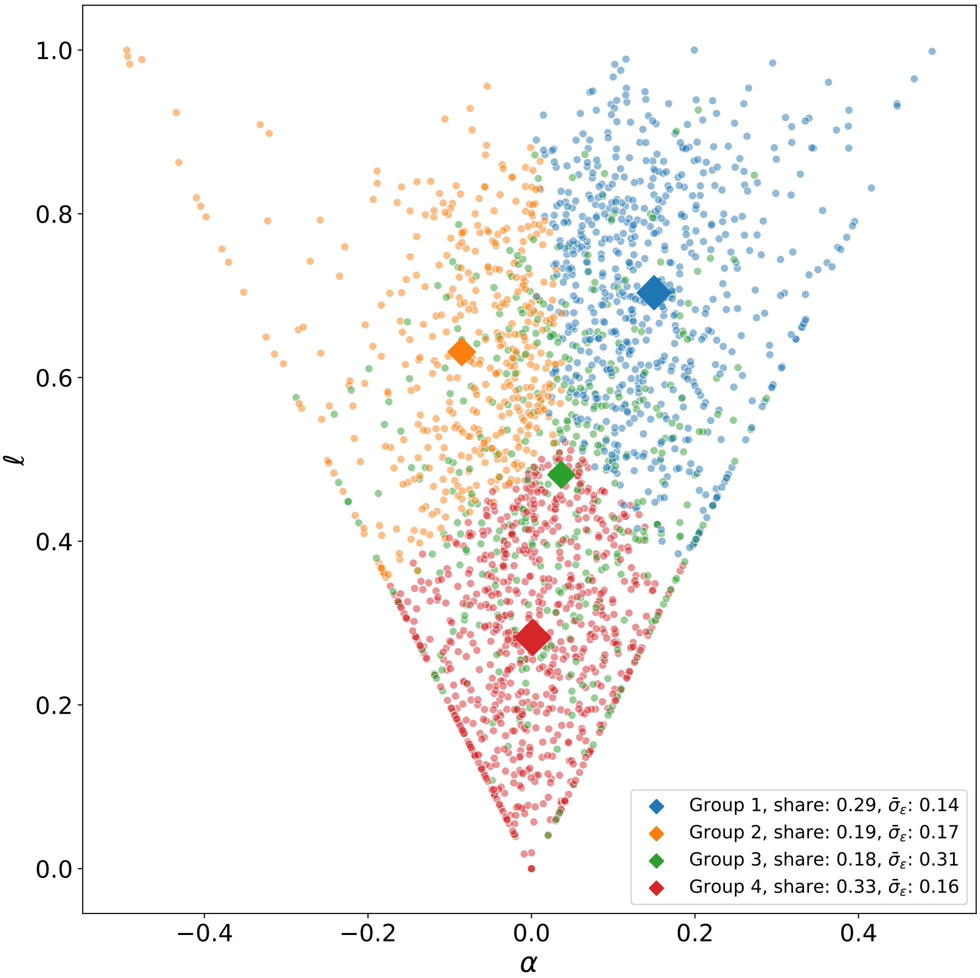


Figure 2.6. Summarising heterogeneity in ambiguity profiles with K=4 discrete groups

Notes: The small dots depict individual ambiguity profiles consisting of the aversion parameter α and the likelihood insensitivity parameter ℓ. The large diamonds are group centres resulting from clustering individuals with the k-means algorithm on the parameters α, ℓ and σ_ε. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects.

behaves inconsistently at comparably small rate ($\sigma = 0.17$). Unlike group 1 individuals, however, subjects in this group are not averse to the ambiguity that they perceive with α ranging from -0.5 to 0. The orange line in Figure 2.7 has a similar slope as the blue line of group 1 due to the comparable ℓ, but is shifted up, crossing the 45-degree line only past the subjective probability of 0.6. This means that individuals in this group exhibit ambiguity seeking behaviour on average, and only become averse to bets on the AEX compared to bets on equally likely lotteries for a high subjective probability of the former. In line with this tendency, the value group

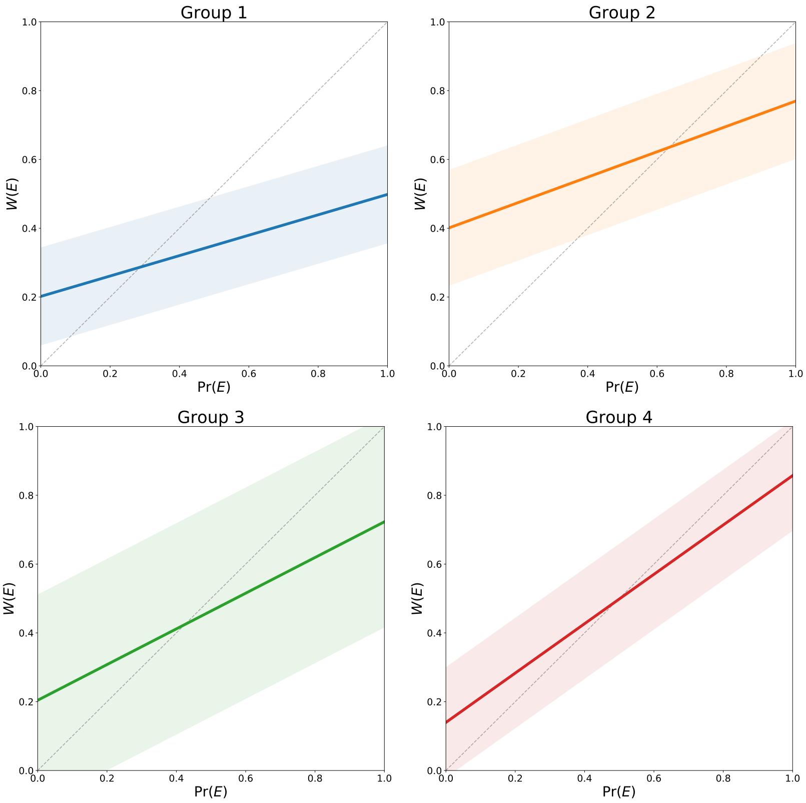


Figure 2.7. Decision weights as a function of subjective probabilities, by group (K=4)

Notes: The figure plots the estimated source functions, i.e. the lines $W(E) = \frac{\ell}{2} - \alpha + (1 - \ell) \cdot \Pr(E)$ for the group-average values of α and ℓ. The vertical difference to the 45-degree line measures the extent of ambiguity seeking w.r.t. gains from events whose source of uncertainty is the future development of the AEX. The shaded area around the lines has bandwidth σ_ε, which visualises the imprecision with which observed matching probabilities measure decision events. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being completed quicker than 85% of subjects.

2's financial assets (median) is 73% higher than for individuals of group 1, they tend to be less risk-averse, and there is a more even gender split. Table 2.6 shows that almost no characteristics predict an individual is more likely to belong to group 1 in the multinomial logistic regression. This is because their characteristics are close to the average of the sample pool.

Group 3: Decisions less consistent with the model, ambiguity parameters not meaningful. 18 % of individuals are characterised by less consistent choices, with σ_ϵ almost doubling compared to the other groups. The green dots in Figure 2.6 are much more spread out, indicating that this group does not form a compact cluster in (α, ℓ) space. A high σ_ϵ can come about through erratic behaviour or because a choice model other than the neoadditive specification we estimated would be appropriate. In line with the former interpretation, individuals in group 3 are characterised by

Table 2.5. Individual characteristics of groups (K=4)

	Group = 1	Group = 2	Group = 3	Group = 4
share	0.29	0.19	0.18	0.33
α	0.15	-0.09	0.04	0.00
ℓ	0.70	0.63	0.48	0.28
σ_ε	0.14	0.17	0.31	0.16
Education: University	0.11	0.11	0.06	0.21
Age	55.74	59.55	64.64	53.92
Female	0.58	0.53	0.47	0.42
Income (thousands)	1.63	1.56	1.47	1.75
Financial assets (thousands)	6.47	10.88	8.82	14.71
Risk aversion index	0.13	-0.04	-0.01	-0.08
Numeracy index	-0.14	0.01	-0.69	0.51
Judged hist. freq: positive return	0.48	0.51	0.48	0.59
Judged hist. freq: response error	0.62	0.60	0.76	0.41
Judged hist. freqs: mean absolute deviation	0.19	0.18	0.21	0.19
Optimism	-0.11	0.07	-0.18	0.15

Notes: The first row shows the share of individuals classified to a given group. For each group, the mean of several variables are shown. For income and total assets, the median is reported instead. The variables for risk aversion, numeracy and optimism are standard normalized. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being completed quicker than 85% of subjects.

a much lower numeracy than those of other groups. They are less educated, made more response errors when judging historical stock returns and their judgements differed the most from actual empirical frequencies. They are on average substantially older than individuals in other groups, and more likely to be men.

Group 4: Low likelihood insensitivity, ambiguity neutral. The remaining third of individuals in our sample shows behaviour close to expected utility maximisation. This group is much less insensitive to changes in subjective probabilities, or equivalently, perceives much less ambiguity, with ℓ only 0.26 on average. There is neither a preponderance of ambiguity aversion nor ambiguity seeking, with the mean value of α equal to 0. Individuals of group 4 do not differ from those of group 1 and 2 in terms of how consistent their decisions are with the model. Figure 2.7 shows that the estimated source function is close to the 45-degree line that characterises expected utility maximisation - decision weights are within one standard deviation σ_ϵ of it over the full range. Individuals of group 4 are more likely to be men and are the youngest on average amongst all four groups, although not much younger than those of group 1. In terms of education, numeracy, as well as the value of financial assets they hold, they score by far the highest. They are also the least risk-averse. Table 2.6 shows that numeracy strongly predicts membership of group 4 conditional on everything else. This is in line with expected utility maximisation being a benchmark of rationality, from which individuals in group 4 fall short the least. Similarly, group 4 individuals stand out for accurately believing that AEX returns were positive around 60% of the time in the past. This optimism is also present in terms of a personality measure.

Our analysis shows that taking into account interdependencies of the three parameters is important; the variance of errors renders the other two parameters less reliable and the magnitude of α is constrained by ℓ. To compare our findings to the existing literature, which does not take interdependencies into account, we also regress parameter estimates on characteristics in table 2.B.7. The patterns are broadly in line with the ones just discussed: α is negatively related to age, financial assets, and numeracy while a higher ℓ is associated with being female, as well as lower education, financial assets, and numeracy. Risk aversion is positively related to both indices once we exclude the high error individuals (group 3), which attenuate relationships in regressions.

Earlier studies on the determinants of ambiguity attitudes report relatively weak connections to demographic variables (Haridon, Vieider, Aycinena, Bandur, Belianin, et al., 2018) and differ in what connections they find. This is likely because they study ambiguity parameters in different settings, and subject pools of varying demographics are used. As our group-based analysis indicates, the second factor can make a sizable difference. One of our key findings, that ℓ is negatively related to both education and numeracy, is in line with Dimmock, Kouwenberg, and Wakker (2015) and Anantanasuwong, Kouwenberg, Mitchell, and Peijnenberg (2019) while Dimmock, Kouwenberg, Mitchell, and Peijnenburg (2015) find a positive relation. There are also opposing findings for the relations of risk aversion and ambiguity attitudes (compare Dimmock, Kouwenberg, Mitchell, and Peijnenburg, 2015; Dimmock, Kouwenberg, and Wakker, 2015; Delavande, Ganguli, and Mengel, 2019). Our results suggest a positive relation to both indices. Contrary to our findings, Butler, Guiso, and Jappelli (2014) find a positive association between wealth and ambiguity aversion.

Table 2.6. Predictors of groups, marginal effects (K=4)

	Group = 1	Group = 2	Group = 3	Group = 4
Age: $\in (35, 50]$	-0.02	0.00	0.02	-0.01
	(0.05)	(0.05)	(0.05)	(0.05)
Age: $\in (50, 65]$	-0.05	0.06	0.04	-0.04
	(0.05)	(0.05)	(0.05)	(0.05)
Age: ≥ 65	-0.08	0.07	0.13**	-0.12**
	(0.05)	(0.05)	(0.05)	(0.05)
Education: Junior college	0.03	-0.01	-0.01	-0.00
	(0.03)	(0.03)	(0.02)	(0.03)
Education: College	0.04	-0.05*	-0.01	0.02
	(0.03)	(0.03)	(0.03)	(0.03)
Education: University	0.01	-0.07	0.01	0.05
	(0.04)	(0.04)	(0.04)	(0.04)
Income: $\in (1.1, 1.6]$	0.05	-0.02	-0.03	-0.01
	(0.03)	(0.03)	(0.03)	(0.03)
Income: $\in (1.6, 2.2]$	0.09***	-0.04	-0.04	-0.01
	(0.03)	(0.03)	(0.03)	(0.03)
Income: ≥ 2.2	0.05	-0.03	-0.01	-0.02
	(0.04)	(0.03)	(0.03)	(0.04)
Financial assets: $\in (1.8, 11.2]$	-0.04	0.01	-0.03	0.05
	(0.03)	(0.03)	(0.03)	(0.04)
Financial assets: $\in (11.2, 32]$	-0.09**	-0.04	0.04	0.09**
	(0.04)	(0.03)	(0.03)	(0.04)
Financial assets: ≥ 32	-0.10***	0.00	0.03	0.08**
	(0.04)	(0.03)	(0.03)	(0.04)
Female	0.07***	0.02	-0.07***	-0.02
	(0.02)	(0.02)	(0.02)	(0.02)
Risk aversion index	0.04***	-0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Numeracy index	-0.04**	-0.00	-0.11***	0.15***
	(0.01)	(0.01)	(0.01)	(0.02)
Judged hist. freq: positive return	-0.14***	-0.03	-0.00	0.18***
	(0.04)	(0.04)	(0.04)	(0.04)
Judged hist. freq: response error	-0.03	0.01	0.05**	-0.04
	(0.03)	(0.02)	(0.02)	(0.02)
Judged hist. freqs: mean absolute deviation	-0.22*	-0.23*	0.40***	0.05
	(0.12)	(0.12)	(0.10)	(0.12)
Optimism	-0.02	0.02	-0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
N	1460	1460	1460	1460
Pseudo R^2	0.12	0.12	0.12	0.12

Notes: Multinomial logit regression, robust standard errors in parentheses. For the thresholds of the income and asset quartiles see Table 2.1. Income and financial assets are in thousands, pooled within household and adjusted for household size. The variables for risk aversion, numeracy and optimism are standardised. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being completed quicker than 85% of subjects.

2.5 Conclusion

This study presented a careful analysis of preferences for decision-making under ambiguity. Motivated by a set of stylised facts, we have set up an empirical model for behaviour in our experiment that features three parameters: ambiguity attitudes, the likelihood insensitivity (or perceived level of ambiguity), and the variance of errors. We have structurally estimated these parameters at the individual level.

Our first main contribution is that we have been able to demonstrate substantial within-person stability of ambiguity attitudes. This holds both over a period of two years and across the domains of financial markets and climate change. In particular, preferences for ambiguity show similar properties as preferences for risk when it comes to stability over time. Across our two contexts, ambiguity aversion is completely stable if we adjust for within-person variability that is due to measurement error, and exhibits stability comparable to risk aversion in measurement error unadjusted comparisons. Likelihood insensitivity, on the other hand, is more variable, strengthening its interpretation as the perceived level of ambiguity, which varies across domains if people are differentially informed. We find some evidence in support of this mechanism; controlling for how much ambiguity individuals perceive in the financial domain, whether they characterise themselves as understanding climate change predicts how much ambiguity they perceive in the climate domain. Nevertheless, there is also a substantial component of this parameter that is stable across contexts.

Our second main contribution has been to describe the patterns of heterogeneity. This has long been done for decisions under risk, but it has proven particularly challenging for decisions under ambiguity. One reason is that all popular models depend on at least two parameters, which are hard to interpret in isolation using parameter-by-parameter regressions. We have instead employed the *k*-means algorithm to classify individuals into a discrete set of groups. Using four groups, we find that a third of the population comes close to the behaviour subjective expected utility maximisers, almost thirty per cent are very averse to ambiguity while almost twenty per cent seek it. The remaining individuals exhibit erratic behaviour. Individuals of these groups systematically differ in background characteristics with reasonable correlations to ambiguity attitudes.

Our key results depend neither on the specifics of the model we use to estimate ambiguity attitudes, nor on the number of groups we use to analyse their heterogeneity. We also estimate ambiguity attitudes in two alternative ways: A version of our model that relaxes parameter restrictions and keeps only the requirement that rules out set-monotonicity errors, and the indices proposed by Baillon, Bleichrodt, Li, and Wakker (2019). Both yield broadly similar results, though the perceived level of ambiguity displays much less stability over time when estimated with the indices. When we double the number of groups in the k-means algorithm, the key results of what is predictive of ambiguity attitudes remain as before.

It remains to be learned how ambiguity attitudes evolve over periods longer than the two years we investigate. A further important follow-up question is how ambiguity attitudes affect behaviour, in particular investment decisions in the financial domain and political, as well as personal decisions regarding climate change. Our design elicits ambiguity attitudes over gains but to understand how ambiguity affects real-world behaviour, ambiguity attitudes over losses might play an important role as well. We leave these questions for future research.

Appendix 2.A Additional figures

Figure 2.A.1 shows the decision tree we use to elicit the matching probability of one aex event. Suppose for example, a subject answered in the following sequence: LOT, AEX, AEX, AEX. Then we would know that the matching probability lies between 40 % and 50 %. Suppose conversely, a subject answered LOT, LOT, LOT, LOT. Then we would know that the matching probability lies between 0 % and 1 %.

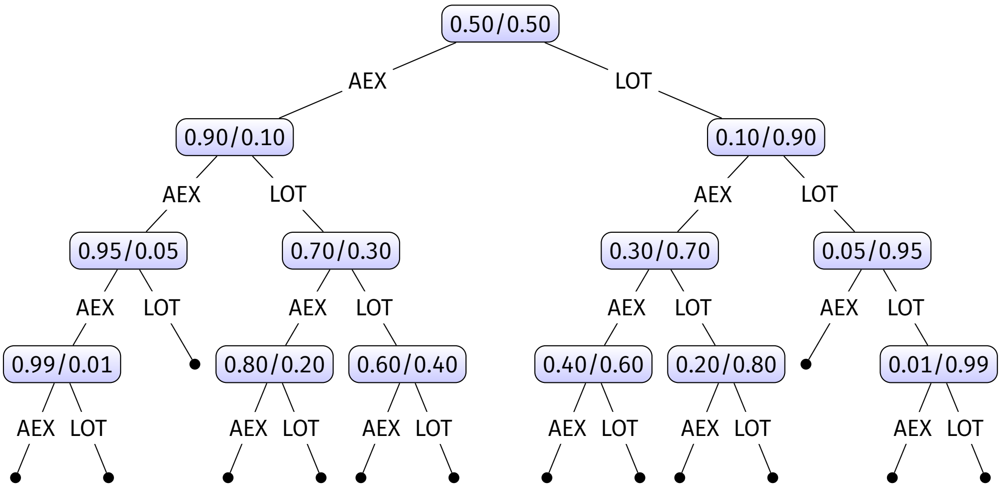


Figure 2.A.1. Iterative sequence of lottery probabilities for one AEX event

Figure 2.A.2 shows the distributions of time taken for the first choice relating to each event, for individuals who used repeating choice patterns for events (always choosing the lottery or always choosing the AEX) and for those who did not.

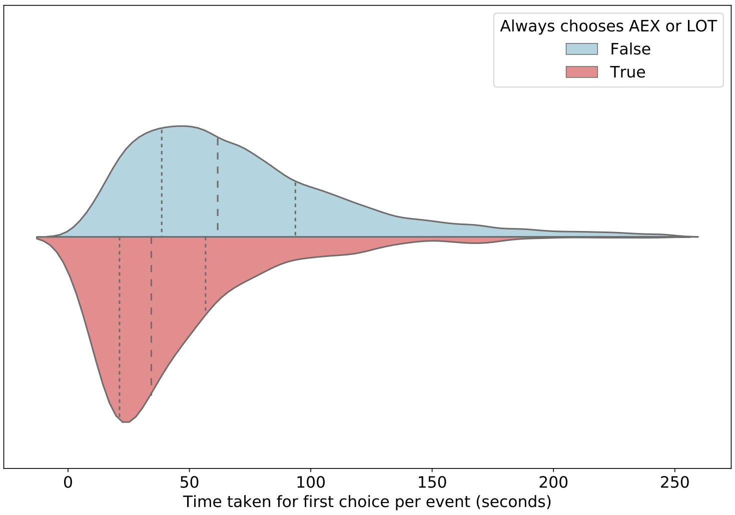


Figure 2.A.2. Time taken for first choice, by choice pattern

Appendix 2.B Additional tables

Table 2.B.1. Matching probabilities by wave

wave	Wave 1	Wave 2	Wave 3	Wave 4	Pooled
$E_0 : r > 0\%$	0.54	0.51	0.52	0.49	0.51
$E_1 : r > 10\%$	0.32	0.35	0.37	0.36	0.35
$E_1^C : r \leq 10\%$	0.58	0.50	0.52	0.52	0.53
$E_2 : r < -5\%$	0.44	0.35	0.34	0.36	0.37
$E_2^C : r \geq -5\%$	0.50	0.54	0.56	0.56	0.54
$E_3 : -5\% \leq r \leq 10\%$	0.58	0.55	0.58	0.58	0.57
$E_3^C : (r < -5\%) \cup (r > 10\%)$	0.41	0.41	0.41	0.41	0.41

Notes: Events were asked about in this order: $E_0 - E_1 - E_2 - E_3 - E_1^C - E_2^C - E_3^C$. Mean of the matching probabilities of the seven events. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being completed quicker than 85% of subjects.

Table 2.B.2. Relation of risk aversion and numeracy with characteristics

	Risk aversion index	Numeracy index
Intercept	-0.37***	-0.28***
	(0.11)	(0.11)
Age: $\in (35, 50]$	0.24**	-0.23**
	(0.10)	(0.09)
Age: $\in (50, 65]$	0.32***	-0.24***
	(0.09)	(0.09)
Age: ≥ 65	0.45***	-0.52***
	(0.10)	(0.09)
Female	0.28***	-0.36***
	(0.05)	(0.04)
Education: Junior college	-0.01	0.19***
	(0.07)	(0.06)
Education: College	0.01	0.42***
	(0.06)	(0.06)
Education: University	-0.15**	0.66***
	(0.07)	(0.06)
Income: $\in (1.1, 1.6]$	-0.07	0.07
	(0.07)	(0.07)
Income: $\in (1.6, 2.2]$	-0.04	0.12*
	(0.08)	(0.06)
Income: ≥ 2.2	-0.20***	0.14**
	(0.08)	(0.06)
Financial assets: $\in (1.8, 11.2]$	-0.08	0.53***
	(0.07)	(0.07)
Financial assets: $\in (11.2, 32]$	0.03	0.72***
	(0.08)	(0.07)
Financial assets: ≥ 32	-0.03	0.78***
	(0.08)	(0.07)
N	1614	1614
R^2	0.049	0.291

Notes: Income and financial assets are in thousands, pooled within household and adjusted for household size. OLS regression, robust standard errors in parentheses.

Table 2.B.3. Subset violations by superset-subset pair

	Subset violations	Δ Judged hist. frequencies
$E_0 \supseteq E_1$	0.09	0.21
$E_1^C \supseteq E_2$	0.10	0.47
$E_1^C \supseteq E_3$	0.22	0.22
$E_2^C \supseteq E_0$	0.19	0.26
$E_2^C \supseteq E_1$	0.10	0.47
$E_2^C \supseteq E_3$	0.20	0.31
$E_3^C \supseteq E_1$	0.16	0.22
$E_3^C \supseteq E_2$	0.18	0.31
Any Violation	0.57	-

Notes: The share of subjects that violate the set-monotonicity conditions for each pair of events is reported in column 1. Set-monotonicity is violated if the interval of the elicited matching probability of the subset is strictly larger than the interval of the superset. The last row shows the share of subjects with at least one error in a given wave. Column 2 shows the difference in the historical frequencies of the respective events.

Table 2.B.4. Relation between subset violations and judged historical frequencies of events

	Superset-Subset Error Rate		
Intercept	0.293***	0.159***	0.075***
	(0.004)	(0.005)	(0.006)
|Jud. freq. superset - Jud. freq. subset| (10 pp)	-0.013***	-0.006***	-0.006***
	(0.001)	(0.001)	(0.001)
Superset - Subset fixed effect	No	Yes	Yes
Individual fixed effect	No	No	Yes
N	15632	15632	15632
R^2	0.02	0.09	0.33

Notes: OLS regressions, robust standard errors in parentheses. The outcomes are individual error rates across waves for all superset-subset event pairs. Standard errors are clustered at the individual level. The regressor is the distance in judged historical frequencies for the events of a superset-subset pair, with unit ten percentage points.

Table 2.B.5. Matching probabilities for temperature questions

	Mean	Std. dev.	$q_{0.1}$	$q_{0.5}$	$q_{0.9}$	Empirical Frequency
$E_0 : \Delta T > 0°C$	0.53	0.27	0.15	0.55	0.92	0.55
$E_1 : \Delta T > 1°C$	0.45	0.27	0.08	0.45	0.92	0.26
$E_1^C : \Delta T \leq 1°C$	0.53	0.28	0.15	0.55	0.92	0.74
$E_2 : \Delta T < -0.5°C$	0.40	0.27	0.03	0.35	0.85	0.26
$E_2^C : \Delta T \geq -0.5°C$	0.50	0.29	0.08	0.45	0.92	0.74
$E_3 : -0.5°C \leq \Delta T \leq 1°C$	0.51	0.28	0.15	0.45	0.92	0.48
$E_3^C : (\Delta T < -0.5°C) \cup (\Delta T > 1°C)$	0.47	0.27	0.08	0.45	0.92	0.52

Notes: Events were elicited in the order $E_0 - E_1 - E_2 - E_3 - E_1^C - E_2^C - E_3^C$. Summary statistics for the matching probabilities of the seven events are shown. The last column shows the empirical frequencies (starting from 1990, own calculation)

Table 2.B.6. Relation between estimated parameter changes and characteristics

	Δ Ambiguity aversion (α)	Δ Perc. level of ambiguity (ℓ)	Δ Model error (σ_ε)
Wave 2	0.02*	0.02	-0.02***
	(0.01)	(0.02)	(0.01)
Wave 3	-0.01	0.03	-0.01*
	(0.01)	(0.02)	(0.01)
Wave 4	-0.01	0.02	0.00
	(0.01)	(0.02)	(0.01)
Age: $\in (35, 50]$	-0.01	-0.02	-0.00
	(0.01)	(0.02)	(0.00)
Age: $\in (50, 65]$	0.00	-0.02	0.00
	(0.01)	(0.02)	(0.00)
Age: ≥ 65	0.01	-0.02	0.01**
	(0.01)	(0.02)	(0.00)
Female	-0.01***	0.01	-0.00
	(0.00)	(0.01)	(0.00)
Education: Junior college	0.01	-0.01	-0.00
	(0.01)	(0.01)	(0.00)
Education: College	0.01*	-0.01	-0.00
	(0.01)	(0.01)	(0.00)
Education: University	0.01	-0.01	-0.01*
	(0.01)	(0.01)	(0.00)
Income: $\in (1.1, 1.6]$	-0.00	0.00	0.00
	(0.01)	(0.01)	(0.00)
Income: $\in (1.6, 2.2]$	-0.01	-0.00	0.00
	(0.01)	(0.01)	(0.00)
Income: ≥ 2.2	0.00	0.00	0.00
	(0.01)	(0.01)	(0.00)
Financial assets: $\in (1.8, 11.2]$	0.01	0.00	-0.00
	(0.01)	(0.01)	(0.00)
Financial assets: $\in (11.2, 32]$	-0.00	-0.00	-0.00
	(0.01)	(0.01)	(0.00)
Financial assets: ≥ 32	0.01	-0.00	-0.01
	(0.01)	(0.01)	(0.00)
Risk aversion index	0.00*	0.00	-0.00
	(0.00)	(0.00)	(0.00)
Numeracy index	-0.01**	-0.00	-0.01***
	(0.00)	(0.00)	(0.00)
N	4181	4181	4181
R^2	0.009	0.001	0.015

Notes: Income and financial assets are in thousands, pooled within household and adjusted for household size. OLS regression, robust standard errors in parentheses. Outcomes are within-subject changes in estimated parameters across consecutive waves.

Table 2.B.7. Relation between estimated parameters and characteristics

	α	α	ℓ	ℓ	σ_ε
Intercept	0.06***	0.05***	0.50***	0.52***	0.17***
	(0.01)	(0.02)	(0.03)	(0.03)	(0.01)
Age: $\in (35, 50]$	-0.01	-0.01	-0.00	0.00	0.02**
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
Age: $\in (50, 65]$	-0.02*	-0.03**	0.01	0.01	0.03***
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
Age: ≥ 65	-0.03**	-0.03**	0.02	0.03	0.05***
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
Female	0.00	0.01	0.03***	0.03***	-0.02***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
Education: Junior college	-0.00	0.00	0.02	0.01	-0.00
	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)
Education: College	-0.01	-0.00	-0.03*	-0.02	-0.01
	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)
Education: University	-0.01	-0.01	-0.05**	-0.04*	-0.00
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
Income: $\in (1.1, 1.6]$	0.01	0.01	0.01	0.01	-0.01*
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
Income: $\in (1.6, 2.2]$	0.01	0.02*	0.02	0.02	-0.01*
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
Income: ≥ 2.2	-0.00	0.00	0.02	0.01	-0.01*
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
Financial assets: $\in (1.8, 11.2]$	-0.00	-0.01	-0.01	-0.02	-0.00
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
Financial assets: $\in (11.2, 32]$	-0.01	-0.02*	-0.04**	-0.04**	0.00
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
Financial assets: ≥ 32	-0.02*	-0.02**	-0.05***	-0.05***	0.00
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
Risk aversion index	0.00	0.01*	0.01	0.01**	-0.01***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)
Numeracy index	-0.01**	-0.01*	-0.04***	-0.07***	-0.03***
	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)
High σ_ε excluded	No	Yes	No	Yes	No
N	1614	1318	1614	1318	1614
R^2	0.024	0.028	0.084	0.139	0.204

Notes: Income and financial assets are in thousands, pooled within household and adjusted for household size. OLS regressions of the parameters of the pooled model on several individual characteristics. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects. In column 2 and column 4, the individuals of the high error group (based on k-means) are excluded. Robust standard errors in parentheses.

Table 2.B.8. Dependence of parameters relating to temperature uncertainty on parameters relating to uncertainty about the AEX

Parameter	α		ℓ		σ_ε	
Model	OLS	2SLS	OLS	2SLS	OLS	2SLS
Intercept	0.05*	0.00	0.44***	0.20***	0.01	-0.03
	(0.03)	(0.03)	(0.06)	(0.07)	(0.02)	(0.02)
AEX param	0.70***	1.00***	0.34***	0.61***	0.48***	1.06***
	(0.03)	(0.09)	(0.03)	(0.06)	(0.03)	(0.11)
Age: $\in (35, 50]$	-0.00	0.00	0.07**	0.10***	0.01	0.01
	(0.01)	(0.01)	(0.03)	(0.04)	(0.01)	(0.01)
Age: $\in (50, 65]$	-0.01	-0.01	0.07**	0.10***	0.01	0.00
	(0.01)	(0.02)	(0.03)	(0.04)	(0.01)	(0.01)
Age: ≥ 65	-0.01	-0.01	0.08**	0.09**	0.01	-0.02
	(0.01)	(0.01)	(0.03)	(0.04)	(0.01)	(0.01)
Education: Junior college	0.00	0.01	0.02	0.03	0.00	-0.01
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Education: College	-0.01	-0.00	0.01	0.02	-0.00	-0.00
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Education: University	-0.01	0.00	-0.05**	-0.03	0.01	0.01
	(0.01)	(0.01)	(0.03)	(0.03)	(0.01)	(0.01)
Income: $\in (1.1, 1.6]$	-0.01	-0.01	-0.01	-0.00	0.00	0.01
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Income: $\in (1.6, 2.2]$	-0.01	-0.01	-0.02	-0.02	-0.00	0.00
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Income: ≥ 2.2	0.00	0.00	-0.01	-0.00	0.00	0.00
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Financial assets: $\in (1.8, 11.2]$	-0.01	0.00	-0.02	-0.01	0.01	0.02*
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Financial assets: $\in (11.2, 32]$	-0.00	0.01	-0.03	-0.00	0.01	-0.00
	(0.01)	(0.01)	(0.02)	(0.03)	(0.01)	(0.01)
Financial assets: ≥ 32	0.01	0.02	-0.04	-0.02	0.01	0.01
	(0.01)	(0.01)	(0.02)	(0.03)	(0.01)	(0.01)
Female	-0.02**	-0.01	0.00	0.00	0.00	0.01**
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Risk aversion index	-0.01*	-0.01**	-0.00	-0.01	-0.00	0.00
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
Numeracy index	0.00	0.01	0.01	0.03**	-0.01***	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Judged hist. freq: positive return	-0.03**	-0.02	0.01	0.05*	0.00	0.01
	(0.01)	(0.02)	(0.03)	(0.03)	(0.01)	(0.01)
Judged hist. freq: response error	0.01*	0.02*	0.02	0.02	0.00	-0.00
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Judged hist. freqs: mean absolute deviation	-0.02	-0.00	-0.06	-0.01	0.02	-0.02
	(0.04)	(0.05)	(0.07)	(0.08)	(0.03)	(0.03)
Optimism	-0.00	0.00	0.01	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
Underst. c.c.	-0.01**	-0.01**	-0.02**	-0.02**	0.01**	0.01***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
Threat. by c.c.	0.00	0.00	-0.00	0.01	0.00	-0.00
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
N	1297	1186	1297	1186	1297	1186
R^2	0.416	-	0.170	-	0.236	-
1st st. F	-	79.8	-	308.4	-	134.9

Notes: Outcomes are estimated parameters in the temperature domain in the 4th wave, regressors are

Appendix 2.C Relaxing parameter restrictions

We restimate our model but keep only the constraint that $\tau_2 > 0$ and the probability constraints, which means that the only behaviour ruled out in the deterministic part of the model are set-monotonicity violations. We calculate the area between the 45 degree line and $\min(\max(\tau_0 + \tau_1 \Pr(E), 0), 1)$ to obtain α, and 1 minus the average slope of $\min(\max(\tau_0 + \tau_1 \Pr(E), 0), 1)$ over the range $Pr(E) \in [0.05, 0.95]$ to obtain ℓ.

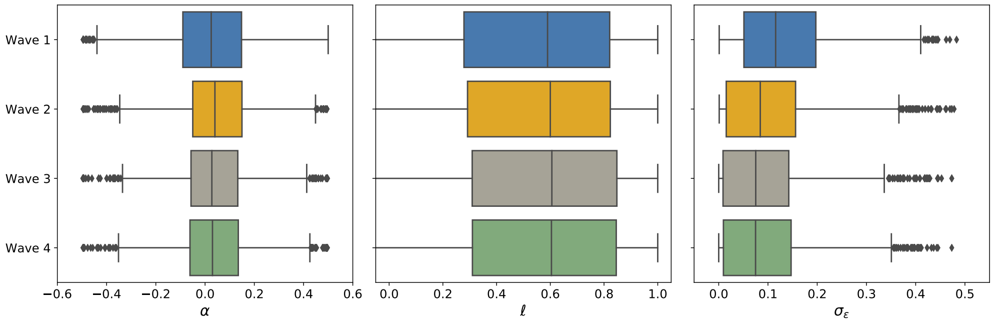


Figure 2.C.1. Distributions of estimated parameters, wave by wave

Notes: Sample restrictions: Observations with regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects

Table 2.C.1. Across wave correlations of estimated parameters

	$\hat{\rho}_{1,2}$	$\hat{\rho}_{1,3}$	$\hat{\rho}_{1,4}$	$\hat{\rho}_{2,3}$	$\hat{\rho}_{2,4}$	$\hat{\rho}_{3,4}$	Average $\hat{\rho}$
α	0.25	0.22	0.20	0.27	0.22	0.33	0.25
ℓ	0.22	0.20	0.24	0.32	0.33	0.39	0.29
σ_ε	0.13	0.17	0.20	0.27	0.28	0.32	0.23

Notes: Table shows Pearson correlations between parameter estimates across waves, with subscripts indicating the waves. Parameter estimates are obtained by the model described in Section 2.3.2 but removing parameter restrictions except $\tau_2 > 0$. The model is estimated separately for each survey wave and individual. Sample restrictions: Observations with regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects.

Table 2.C.2. Dependence of parameters relating to temperature uncertainty on parameters relating to uncertainty about the AEX

Parameter		α			ℓ			σ_ε	
Model	OLS	OLS	2SLS	OLS	OLS	2SLS	OLS	OLS	2SLS
Intercept	-0.01**	0.05*	0.01	0.41***	0.42***	0.16**	0.06***	0.01	-0.04
	(0.00)	(0.03)	(0.03)	(0.02)	(0.06)	(0.08)	(0.00)	(0.02)	(0.02)
AEX param	0.70***	0.70***	1.00***	0.36***	0.35***	0.65***	0.49***	0.46***	1.12***
	(0.03)	(0.03)	(0.09)	(0.03)	(0.03)	(0.07)	(0.03)	(0.03)	(0.12)
Underst. c.c.		-0.01**	-0.01**		-0.02**	-0.02**		0.01**	0.01***
		(0.00)	(0.00)		(0.01)	(0.01)		(0.00)	(0.00)
Threat. by c.c.		0.00	0.00		-0.00	0.01		0.00	0.00
		(0.00)	(0.00)		(0.01)	(0.01)		(0.00)	(0.00)
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
N	1297	1297	1186	1297	1297	1186	1297	1297	1186
R^2	0.400	0.414	-	0.139	0.159	-	0.202	0.223	-
1st st. F	-	-	79.9	-	-	256.6	-	-	81.6

Notes: Sample restrictions: Observations with regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects. Robust standard errors in parentheses. Outcomes are estimated parameters in the temperature domain in the 4th wave, regressors are estimated parameters in the AEX domain in the 4th wave. Two stage least squares models use estimated parameters from the previous three waves as instruments. Controls are age, gender, education, income and assets dummies, risk aversion, basic, financial and probability numeracy and indicators of self-assessed understanding and perceived threat of climate change with a 5 and 6 point scale respectively.

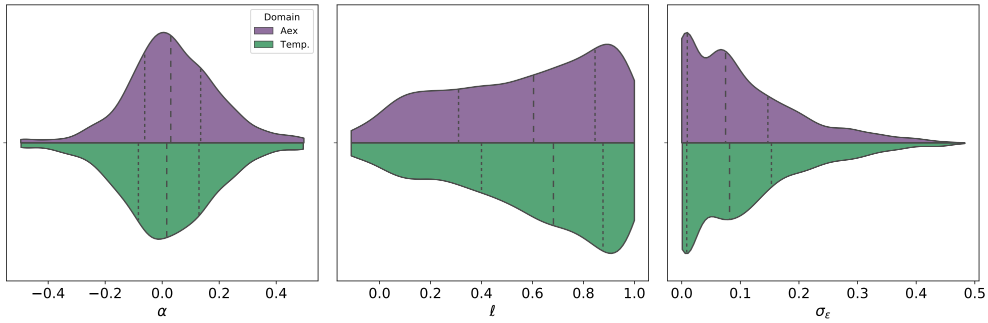


Figure 2.C.2. Distributions of estimated parameters, AEX v Temperature domains

Notes: Sample restrictions: Observations with regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects

Table 2.C.3. Individual characteristics of groups (K=4)

	Group = 1	Group = 2	Group = 3	Group = 4
share	0.29	0.21	0.18	0.32
α	0.15	-0.07	0.03	0.00
ℓ	0.70	0.62	0.49	0.26
σ_ε	0.14	0.16	0.31	0.16
Education: University	0.11	0.13	0.05	0.20
Age	55.51	58.05	65.07	54.79
Female	0.58	0.53	0.47	0.41
Income (thousands)	1.62	1.58	1.47	1.75
Financial assets (thousands)	6.32	10.89	10.00	14.71
Risk aversion index	0.12	-0.05	-0.02	-0.07
Numeracy index	-0.12	0.05	-0.72	0.48
Judged hist. freq: positive return	0.48	0.53	0.47	0.58
Judged hist. freq: response error	0.62	0.57	0.77	0.43
Judged hist. freqs: mean absolute deviation	0.18	0.18	0.22	0.19
Optimism	-0.11	0.06	-0.18	0.16

Notes: The first row shows the share of individuals classified to a given group. For each group, the mean of several variables are shown. For income and total assets, the median is reported instead. The variables for risk aversion, numeracy and optimism are standard normalized. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being completed quicker than 85% of subjects.

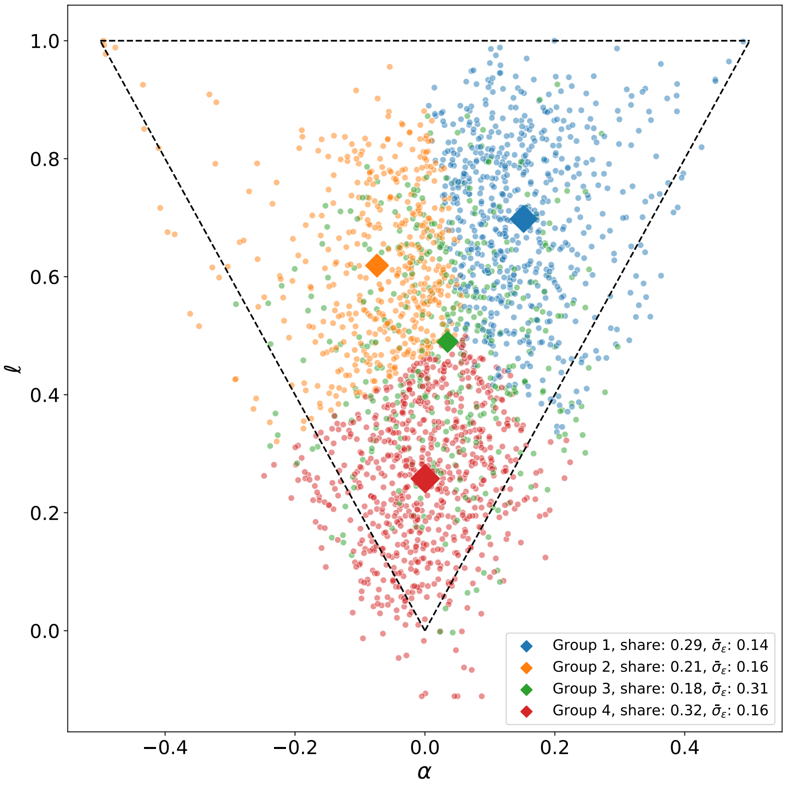


Figure 2.C.3. Summarising heterogeneity in ambiguity profiles with K=4 discrete groups

Notes: The small dots depict individual ambiguity profiles consisting of the aversion parameter α and the likelihood insensitivity parameter ℓ. The large diamonds are group centres resulting from clustering individuals with the k-means algorithm on the parameters α, ℓ and σ_{ε}. We dashed black triangle shows the region into which we constrain estimates in our main model. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects.

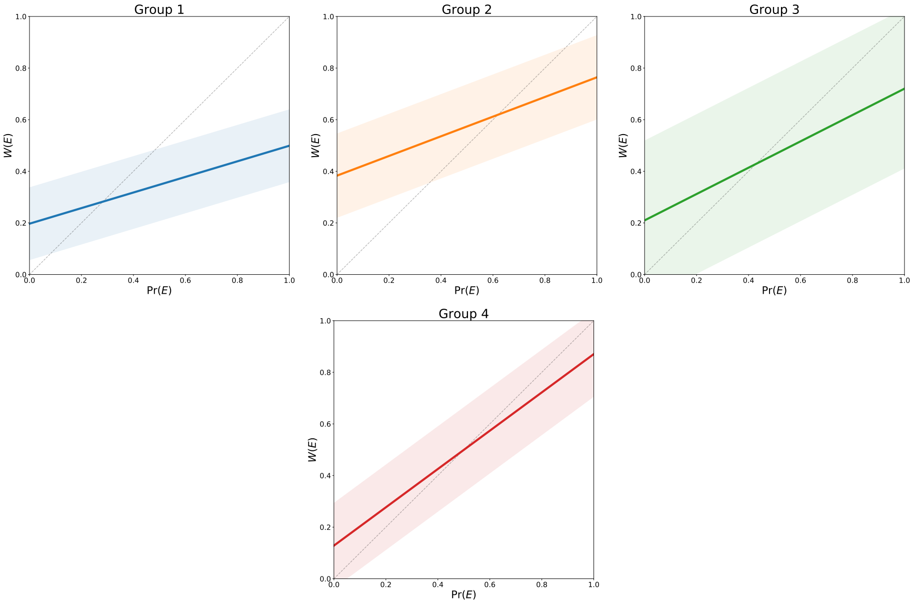


Figure 2.C.4. Event weights as a function of subjective probabilities, by group (K=4)

Notes: Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being completed quicker than 85% of subjects. The figure plots the lines $W(E) = \frac{\ell}{2} - \alpha + (1 - \ell) \cdot \mathrm{Pr}(E)$ for the group-average values of α and ℓ. The vertical difference to the 45 degree line measures the extent of ambiguity seeking w.r.t. gains from events whose source of uncertainty is the future development of the AEX.

Table 2.C.4. Predictors of groups, marginal effects (K=4)

	Group = 1	Group = 2	Group = 3	Group = 4
Age: $\in (35, 50]$	-0.02	-0.03	0.05	-0.00
	(0.05)	(0.05)	(0.06)	(0.05)
Age: $\in (50, 65]$	-0.08	0.02	0.08	-0.02
	(0.05)	(0.05)	(0.06)	(0.05)
Age: ≥ 65	-0.10**	0.01	0.18***	-0.08*
	(0.05)	(0.05)	(0.06)	(0.05)
Education: Junior college	0.02	-0.01	0.00	-0.01
	(0.03)	(0.03)	(0.02)	(0.03)
Education: College	0.04	-0.05	-0.00	0.01
	(0.03)	(0.03)	(0.03)	(0.03)
Education: University	-0.00	-0.03	0.01	0.02
	(0.04)	(0.04)	(0.04)	(0.04)
Income: $\in (1.1, 1.6]$	0.05*	-0.02	-0.02	-0.01
	(0.03)	(0.03)	(0.03)	(0.03)
Income: $\in (1.6, 2.2]$	0.09***	-0.04	-0.03	-0.02
	(0.03)	(0.03)	(0.03)	(0.03)
Income: ≥ 2.2	0.06*	-0.04	-0.01	-0.01
	(0.04)	(0.03)	(0.03)	(0.04)
Financial assets: $\in (1.8, 11.2]$	-0.05	0.04	-0.03	0.03
	(0.03)	(0.03)	(0.03)	(0.04)
Financial assets: $\in (11.2, 32]$	-0.08**	-0.01	0.03	0.06*
	(0.03)	(0.04)	(0.03)	(0.04)
Financial assets: ≥ 32	-0.11***	0.02	0.03	0.07*
	(0.04)	(0.04)	(0.03)	(0.04)
Female	0.07***	0.02	-0.07***	-0.01
	(0.02)	(0.02)	(0.02)	(0.02)
Risk aversion index	0.03***	-0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Numeracy index	-0.03*	-0.01	-0.11***	0.15***
	(0.01)	(0.01)	(0.01)	(0.02)
Judged hist. freq: positive return	-0.14***	-0.01	-0.01	0.16***
	(0.04)	(0.04)	(0.03)	(0.04)
Judged hist. freq: response error	-0.01	-0.00	0.05**	-0.03
	(0.03)	(0.02)	(0.02)	(0.02)
Judged hist. freqs: mean absolute deviation	-0.30**	-0.23*	0.42***	0.11
	(0.12)	(0.12)	(0.10)	(0.12)
Optimism	-0.02*	0.02	-0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
N	1460	1460	1460	1460
Pseudo R^2	0.12	0.12	0.12	0.12

Notes: Income and financial assets are in thousands, pooled within household and adjusted for household size. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being completed quicker than 85% of subjects.

Appendix 2.D Analysis with indices

We estimate ambiguity attitudes using the indices proposed by Baillon, Bleichrodt, Li, and Wakker (2019) except that to maintain comparability with our main results, we do not divide the estimate of the ambiguity aversion parameter α by 2.

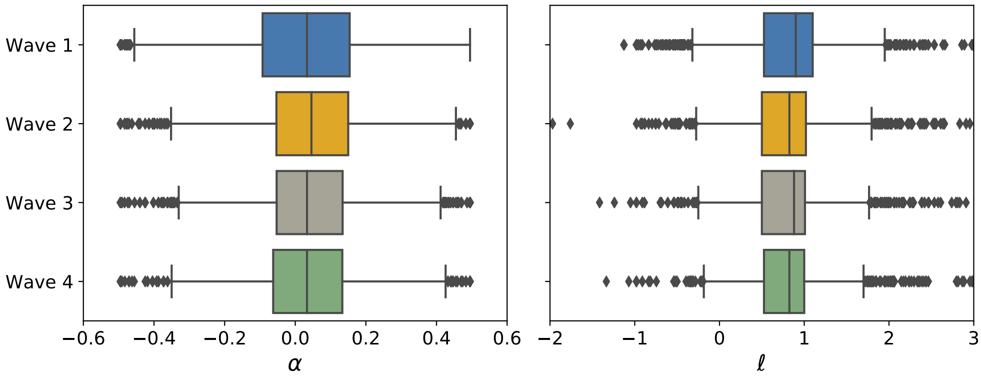


Figure 2.D.1. Distributions of estimated parameters, wave by wave

Notes: Sample restrictions: Observations with regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects. Parameter estimates are the indices proposed by Baillon, Bleichrodt, Li, and Wakker (2019), calculated for each survey wave and individual.

Table 2.D.1. Across wave correlations of estimated parameters

	$\hat{\rho}_{1,2}$	$\hat{\rho}_{1,3}$	$\hat{\rho}_{1,4}$	$\hat{\rho}_{2,3}$	$\hat{\rho}_{2,4}$	$\hat{\rho}_{3,4}$	Average $\hat{\rho}$
α	0.21	0.18	0.24	0.32	0.25	0.20	0.24
ℓ	-0.02	0.02	0.03	0.19	0.15	0.15	0.09

Notes: Sample restrictions: Observations with regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects. Table shows Pearson correlations between parameter estimates across waves, with subscripts indicating the waves. Parameter estimates are the indices proposed by Baillon, Bleichrodt, Li, and Wakker (2019), calculated for each survey wave and individual.

Table 2.D.2. Dependence of parameters relating to temperature uncertainty on parameters relating to uncertainty about the AEX

Parameter		α			ℓ	
Model	OLS	OLS	2SLS	OLS	OLS	2SLS
Intercept	-0.00	0.05*	0.01	0.73***	0.63***	0.06
	(0.00)	(0.03)	(0.03)	(0.03)	(0.10)	(0.23)
AEX param	0.69***	0.68***	1.05***	0.16***	0.14***	0.79***
	(0.03)	(0.03)	(0.10)	(0.04)	(0.04)	(0.23)
Underst. c.c.		-0.01**	-0.01**		-0.01	-0.01
		(0.00)	(0.01)		(0.01)	(0.02)
Threat. by c.c.		0.01	0.01		0.01	0.02
		(0.00)	(0.00)		(0.01)	(0.02)
Controls	No	Yes	Yes	No	Yes	Yes
N	1297	1297	1186	1297	1297	1186
R^2	0.386	0.400	-	0.028	0.053	-
1st st. F	-	-	67.1	-	-	24.4

Notes: Sample restrictions: Observations with regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects. Robust standard errors in parentheses. Outcomes are estimated parameters in the temperature domain in the 4th wave, regressors are estimated parameters in the AEX domain in the 4th wave. Two stage least squares models use estimated parameters from the previous three waves as instruments. Controls are age, gender, education, income and assets dummies, risk aversion, basic, financial and probability numeracy and indicators of self-assessed understanding and perceived threat of climate change with a 5 and 6 point scale respectively. Parameter estimates are the indices proposed by Baillon, Bleichrodt, Li, and Wakker (2019), calculated for each survey wave and individual.

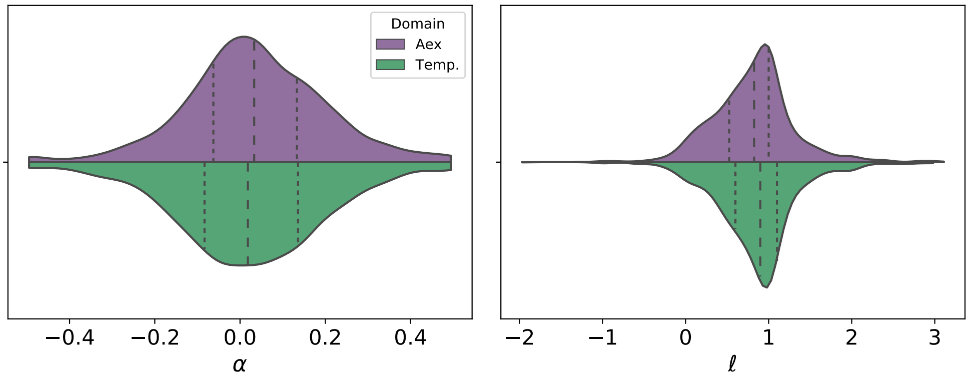


Figure 2.D.2. Distributions of estimated parameters, AEX v Temperature domains

Notes: Sample restrictions: Observations with regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects. Parameter estimates are the indices proposed by Baillon, Bleichrodt, Li, and Wakker (2019), calculated for each survey wave and individual.

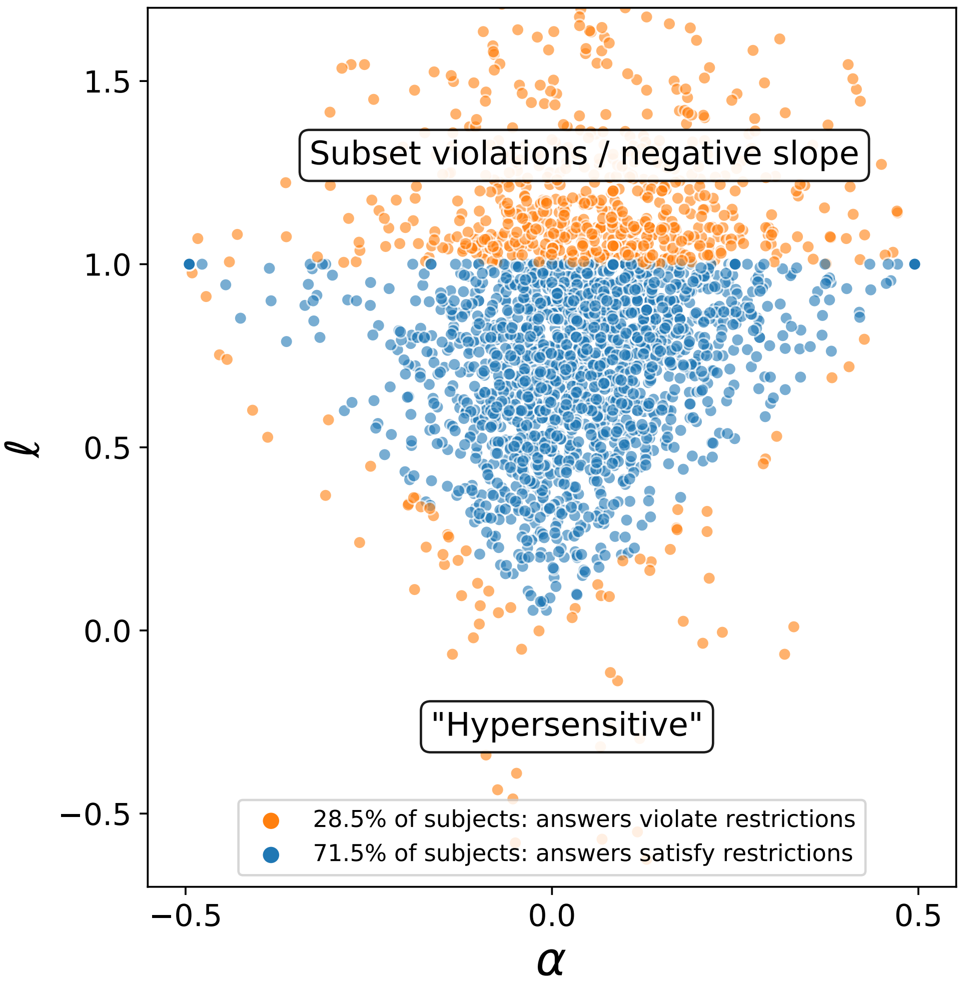


Figure 2.D.3. Summarising heterogeneity in ambiguity profiles, indices

Notes: Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects. Parameter estimates are across-wave averages of the indices proposed by Baillon, Bleichrodt, Li, and Wakker (2019). The blue dots are parameter values that satisfy the restrictions we impose in our main model. Values above can only came about through set-monotonicity errors. Values below indicate hypersensitivity.

Table 2.D.3. Relation between estimated indices and characteristics

	α	ℓ
Intercept	0.06***	0.78***
	(0.01)	(0.03)
Age: $\in (35, 50]$	-0.01	0.03
	(0.01)	(0.03)
Age: $\in (50, 65]$	-0.02*	0.04
	(0.01)	(0.03)
Age: ≥ 65	-0.02**	0.08**
	(0.01)	(0.03)
Female	0.01	-0.00
	(0.01)	(0.01)
Education: Junior college	-0.00	0.03
	(0.01)	(0.02)
Education: College	-0.01	-0.06***
	(0.01)	(0.02)
Education: University	-0.01	-0.07***
	(0.01)	(0.03)
Income: $\in (1.1, 1.6]$	0.01	-0.01
	(0.01)	(0.02)
Income: $\in (1.6, 2.2]$	0.01	0.02
	(0.01)	(0.02)
Income: ≥ 2.2	-0.00	0.01
	(0.01)	(0.02)
Financial assets: $\in (1.8, 11.2]$	-0.00	-0.01
	(0.01)	(0.02)
Financial assets: $\in (11.2, 32]$	-0.01	-0.01
	(0.01)	(0.02)
Financial assets: ≥ 32	-0.02*	-0.02
	(0.01)	(0.02)
Risk aversion index	0.00	-0.00
	(0.00)	(0.01)
Numeracy index	-0.01*	-0.07***
	(0.00)	(0.01)
N	1614	1614
R^2	0.021	0.093

Notes: OLS regressions of the ambiguity indices pooled over all waves on several individual characteristics. Income and financial assets are in thousands, pooled within household and adjusted for household size. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects. Robust standard errors in parentheses.

Appendix 2.E Setting the number of groups to K = 8

We double the number of groups from $K = 4$ to $K = 8$ when allocating individuals into groups with the k-means algorithm and reproduce the analyses of Section 2.4.3.

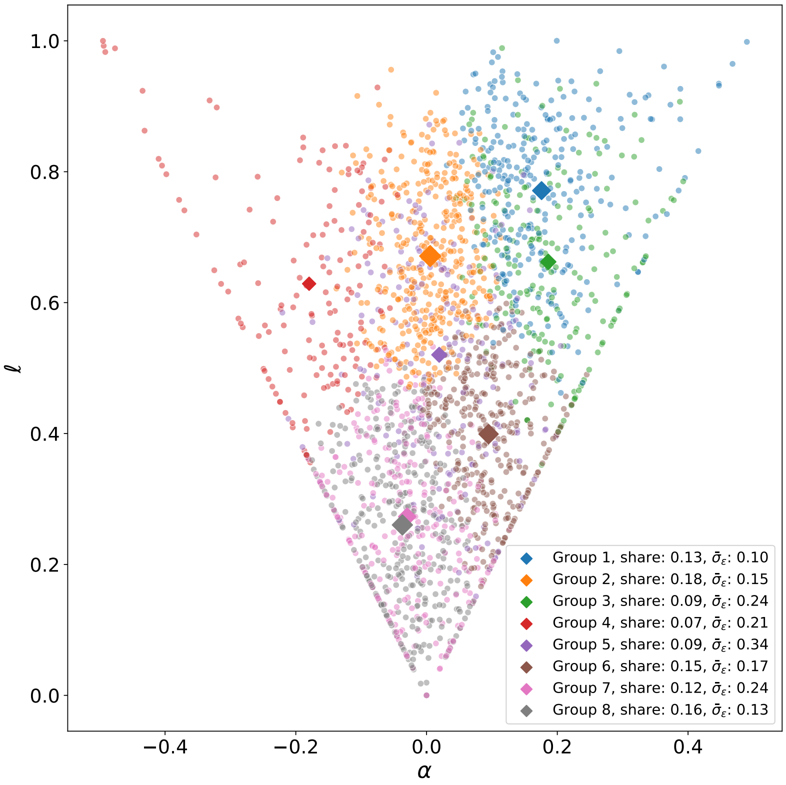


Figure 2.E.1. Summarising heterogeneity in ambiguity profiles with K=8 discrete groups

Notes: The small dots depict individual ambiguity profiles consisting of the aversion parameter α and the likelihood insensitivity parameter ℓ. The large diamonds are group centres resulting from clustering individuals with the k-means algorithm on the parameters α, ℓ and σ_ε. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being entered quicker than 85% of subjects.

Table 2.E.1. Individual characteristics of groups (K=8)

	G = 1	G = 2	G = 3	G = 4	G = 5	G = 6	G = 7	G = 8
share	0.13	0.18	0.09	0.07	0.09	0.15	0.12	0.16
α	0.18	0.01	0.19	-0.18	0.02	0.09	-0.03	-0.04
ℓ	0.77	0.67	0.66	0.63	0.52	0.40	0.27	0.26
σ_ε	0.10	0.15	0.24	0.21	0.34	0.17	0.24	0.13
Education: University	0.10	0.12	0.07	0.10	0.05	0.12	0.15	0.26
Age	54.73	56.37	61.60	63.46	66.57	54.16	60.88	51.61
Female	0.67	0.56	0.47	0.51	0.49	0.48	0.39	0.40
Income (thousands)	1.58	1.61	1.47	1.54	1.40	1.70	1.74	1.76
Financial assets (thousands)	5.07	9.56	6.40	11.76	9.85	11.47	14.68	14.71
Risk aversion index	0.17	-0.02	0.13	0.08	-0.09	0.00	-0.11	-0.10
Numeracy index	-0.16	0.02	-0.52	-0.20	-0.85	0.35	-0.01	0.66
Judged hist. freq: positive return	0.46	0.50	0.45	0.50	0.50	0.55	0.53	0.62
Judged hist. freq: response error	0.64	0.56	0.73	0.67	0.77	0.51	0.59	0.35
Judged hist. freqs: mean absolute deviation	0.19	0.18	0.21	0.19	0.21	0.19	0.21	0.18
Optimism	-0.10	-0.00	-0.22	0.04	-0.23	0.07	0.08	0.18

Notes: The first row shows the share of individuals classified to a given group. For each group, the mean of several variables are shown. For income and total assets, the median is reported instead. The variables for risk aversion, numeracy and optimism are standard normalized. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being completed quicker than 85% of subjects.

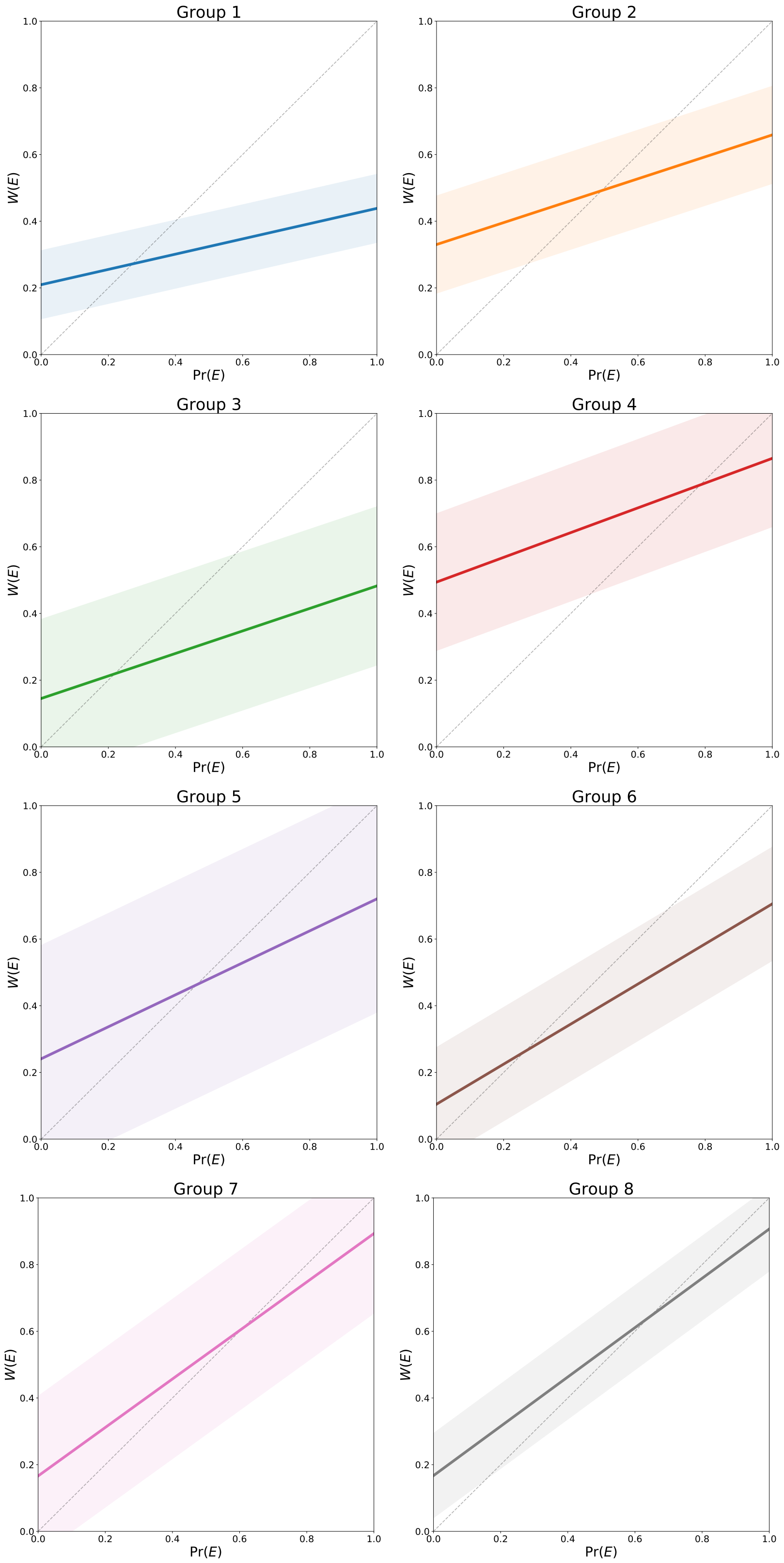


Figure 2.E.2. Event weights as a function of subjective probabilities, by group (K=8)

Notes: The figure plots the estimated source functions, i.e. the lines $W(E) = \frac{\ell}{2} - \alpha + (1-\ell) \cdot \Pr(E)$ for the group-average values of α and ℓ. The vertical difference to the 45 degree line measures the extent of ambiguity seeking w.r.t. gains from events whose source of uncertainty is the future development of the AEX. The shaded area around the lines has bandwith σ_ε, which visualises the imprecision with which observed matching probabilities measure event events. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being completed quicker than 85% of subjects.

Table 2.E.2. Predictors of groups, marginal effects (K=8)

	G = 1	G = 2	G = 3	G = 4	G = 5	G = 6	G = 7	G = 8
Age: $\in (35, 50]$	-0.00	-0.02	-0.05	0.09	0.02	-0.04	0.04	-0.04
	(0.03)	(0.05)	(0.04)	(0.07)	(0.06)	(0.04)	(0.05)	(0.04)
Age: $\in (50, 65]$	-0.02	-0.03	-0.03	0.10	0.04	-0.11***	0.08	-0.03
	(0.03)	(0.05)	(0.04)	(0.07)	(0.05)	(0.04)	(0.05)	(0.04)
Age: ≥ 65	-0.05	-0.03	-0.01	0.12*	0.10*	-0.09**	0.07	-0.11***
	(0.03)	(0.05)	(0.04)	(0.07)	(0.05)	(0.04)	(0.05)	(0.04)
Education: Junior college	0.01	0.02	-0.01	-0.03*	-0.01	-0.01	0.06**	-0.02
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Education: College	0.01	-0.01	-0.00	-0.03*	-0.02	-0.02	0.06**	0.02
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Education: University	0.02	-0.02	-0.02	-0.03	0.01	-0.10***	0.10***	0.04
	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Income: $\in (1.1, 1.6]$	0.03	0.02	0.02	-0.03	-0.02	-0.04	-0.01	0.04
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Income: $\in (1.6, 2.2]$	0.05**	-0.00	0.01	0.00	-0.02	0.01	-0.02	-0.02
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Income: ≥ 2.2	0.03	-0.01	-0.00	-0.00	-0.03	-0.01	0.01	0.01
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Financial assets: $\in (1.8, 11.2]$	-0.04	0.01	-0.03	0.00	-0.03	0.04	-0.03	0.06*
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Financial assets: $\in (11.2, 32]$	-0.05**	-0.01	-0.03	-0.01	0.01	0.01	0.03	0.05
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Financial assets: ≥ 32	-0.08***	-0.02	-0.03	0.01	0.01	0.01	0.03	0.06*
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Female	0.07***	0.03	-0.03*	-0.00	-0.04**	0.02	-0.04*	-0.00
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)
Risk aversion index	0.02***	-0.00	0.00	0.00	-0.01	-0.00	-0.01	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Numeracy index	-0.02*	-0.01	-0.03***	-0.02**	-0.06***	0.05***	-0.04***	0.12***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Judged hist. freq: positive return	-0.08**	-0.10**	-0.04	-0.01	0.02	0.03	0.00	0.18***
	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
Judged hist. freq: response error	-0.01	-0.02	0.01	0.00	0.01	0.01	0.01	-0.02
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)
Judged hist. freqs: mean abs. dev.	-0.06	-0.41***	0.14*	0.01	0.14*	-0.03	0.29***	-0.08
	(0.09)	(0.12)	(0.09)	(0.07)	(0.08)	(0.10)	(0.10)	(0.10)
Optimism	-0.01	0.02	-0.01*	0.01	-0.01	-0.00	0.01	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
N	1460	1460	1460	1460	1460	1460	1460	1460
Pseudo R^2	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10

Notes: Multinomial logit regression. Robust standard errors. For the thresholds of the income and asset quartiles see Table 2.1. Income and financial assets are in thousands, pooled within household and adjusted for household size. The variables for risk aversion, numeracy and optimism are standard normalized. Sample restrictions: Individuals with at least two waves of regular choices. Choices are irregular if they exhibit recurring patterns whilst also being completed quicker than 85% of subjects.

References

Abdellaoui, Mohammed, Aurélien Baillon, Laetitia Placido, and Peter P. Wakker. 2011. "The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation." *American Economic Review* 101 (2): 695–723. [90, 98]

Anantanasuwong, Kanin, Roy Kouwenberg, Olivia S Mitchell, and Kim Peijnenberg. 2019. "Ambiguity Attitudes about Investments: Evidence from the Field." Working Paper. [90, 106, 107, 112]

Baillon, Aurélien, and Han Bleichrodt. 2015. "Testing Ambiguity Models through the Measurement of Probabilities for Gains and Losses." *American Economic Journal: Microeconomics* 7 (2): 77–100. [90]

Baillon, Aurélien, Han Bleichrodt, Umut Keskin, Olivier l'Haridon, and Chen Li. 2018. "The Effect of Learning on Ambiguity Attitudes." *Management Science* 64 (5): 2181–98. [90, 100, 107]

Baillon, Aurélien, Han Bleichrodt, Chen Li, and Peter P Wakker. 2019. "Belief Hedges: Applying Ambiguity Measurements to All Events and All Ambiguity Models," 48. [102, 103, 114, 130–133]

Baillon, Aurelien, Yoram Halevy, and Chen Li. 2014. "Experimental Elicitation of Ambiguity Attitude." Working Paper. [96]

Baillon, Aurélien, Zhenxing Huang, Asli Selim, and Peter P. Wakker. 2018. "Measuring Ambiguity Attitudes for All (Natural) Events." *Econometrica*, [90, 94, 95]

Bardsley, Nicholas. 2000. "Control without Deception: Individual Behaviour in Free-Riding Experiments Revisited." *Experimental Economics* 3 (3): 215–40. [96]

Bianchi, Milo, and Jean-Marc Tallon. 29, 2018. "Ambiguity Preferences and Portfolio Choices: Evidence from the Field." *Management Science*, [90]

Bruine de Bruin, Wändi Bruine de, Baruch Fischhoff, Susan G. Millstein, and Bonnie L. Halpern-Felsher. 2000. "Verbal and Numerical Expressions of Probability: "It's a Fifty–Fifty Chance"." *Organizational Behavior and Human Decision Processes* 81 (1): 115–31. [89]

Butler, Jeffrey V., Luigi Guiso, and Tullio Jappelli. 2014. "The Role of Intuition and Reasoning in Driving Aversion to Risk and Ambiguity." *Theory and Decision* 77 (4): 455–84. [89, 112]

Chateauneuf, Alain, Jürgen Eichberger, and Simon Grant. 1, 2007. "Choice under Uncertainty with the Best and Worst in Mind: Neo-Additive Capacities." *Journal of Economic Theory* 137 (1): 538–67. [89, 99, 100]

Chuang, Yating, and Laura Schechter. 1, 2015. "Stability of Experimental and Survey Measures of Risk, Time, and Social Preferences: A Review and Some New Results." *Journal of Development Economics* 117: 151–70. [91, 104]

Delavande, Adeline, Jayant Ganguli, and Friederike Mengel. 2019. "Measuring Uncertainty Attitudes and Their Impact on Behaviour in General Social Surveys." Working Paper. [90, 112]

Dimmock, Stephen G., Roy Kouwenberg, Olivia S. Mitchell, and Kim Peijnenburg. 1, 2015. "Estimating Ambiguity Preferences and Perceptions in Multiple Prior Models: Evidence from the Field." *Journal of Risk and Uncertainty* 51 (3): 219–44. [89, 112]

Dimmock, Stephen G., Roy Kouwenberg, Olivia S. Mitchell, and Kim Peijnenburg. 2016. "Ambiguity Aversion and Household Portfolio Choice Puzzles: Empirical Evidence." *Journal of Financial Economics* 119 (3): 559–77. [90]

Dimmock, Stephen G., Roy Kouwenberg, and Peter P. Wakker. 2, 2015. "Ambiguity Attitudes in a Large Representative Sample." *Management Science* 62 (5): 1363–80. [89, 112]

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 1, 2011. “Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences.” *Journal of the European Economic Association* 9 (3): 522–50. [93, 107]

Drerup, Tilman, Benjamin Enke, and Hans-Martin von Gaudecker. 1, 2017. “The Precision of Subjective Data and the Explanatory Power of Economic Models.” *Journal of Econometrics*. Measurement Error Models 200 (2): 378–89. [89]

Ellsberg, Daniel. 1961. “Risk, Ambiguity, and the Savage Axioms.” *The Quarterly Journal of Economics* 75 (4): 643–69. [89]

Enke, Benjamin, and Thomas Graeber. 2019. “Cognitive Uncertainty.” Working Paper. [90]

Falk, Armin, Anke Becker, Thomas J. Dohmen, David Huffman, and Uwe Sunde. 2016. “The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences.” *SSRN Electronic Journal*, [92, 93]

Ghirardato, Paolo, and Massimo Marinacci. 1, 2001. “Risk, Ambiguity, and the Separation of Utility and Beliefs.” *Mathematics of Operations Research* 26 (4): 864–90. [89, 98]

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2018. “Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study.” *Journal of Political Economy*, 45. [91]

Haridon, Olivier l’, Ferdinand M. Vieider, Diego Aycinena, Agustinus Bandur, Alexis Belianin, Lubomír Cingl, Amit Kothiyal, and Peter Martinsson. 12, 2018. “Off the Charts: Massive Unexplained Heterogeneity in a Global Study of Ambiguity Attitudes.” *The Review of Economics and Statistics* 100 (4): 664–77. [112]

Hudomiet, Peter, Michael Hurd, and Susann Rohwedder. 2018. *Measuring Probability Numeracy*. RAND Corporation. [93]

Hurd, Michael D. 2009. “Subjective Probabilities in Household Surveys.” *Annual Review of Economics* 1 (1): 543–62. [89]

Johnson, Cathleen A., Aurelien Baillon, Han Bleichrodt, Zhihua Li, Dennie van Dolder, and Peter P. Wakker. 16, 2015. “Prince: An Improved Method for Measuring Incentivized Preferences.” Working Paper. [96]

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji. 2005. “A Smooth Model of Decision Making under Ambiguity.” *Econometrica* 73 (6): 1849–92. [89]

Li, Zhihua, Julia Müller, Peter P. Wakker, and Tong V. Wang. 2018. “The Rich Domain of Ambiguity Explored.” *Management Science* 64 (7): 3227–40. [89, 99]

Manski, Charles F. 2004. “Measuring Expectations.” *Econometrica* 72 (5): 1329–76. [89]

Scheier, Michael F, Charles S Carver, and Michael W Bridges. 1994. “Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test.” *Journal of personality and social psychology* 67 (6): 1063. [94]

Schildberg-Hörisch, Hannah. 2018. “Are Risk Preferences Stable?” *Journal of Economic Perspectives* 32 (2): 135–54. [104]

Steptoe, Andrew, Elizabeth Breeze, James Banks, and James Nazroo. 2013. “Cohort Profile: The English Longitudinal Study of Ageing.” *International journal of epidemiology* 42 (6): 1640–48. [93]

Trautmann, Stefan T., and Gijs van de Kuilen. 18, 2015. “Ambiguity Attitudes.” In *The Wiley Blackwell Handbook of Judgment and Decision Making*. Edited by Gideon Keren and George Wu. Chichester, UK: John Wiley & Sons, Ltd, 89–116. [89, 99]

Van Rooij, Maarten, Annamaria Lusardi, and Rob Alessie. 2011. “Financial Literacy and Stock Market Participation.” *Journal of Financial Economics* 101 (2): 449–72. [93]

Wakker, Peter P. 2010. *Prospect Theory: For Risk and Ambiguity*. Cambridge: Cambridge University Press. [90]

Chapter 3

Can betting odds be turned into informative probabilities of events? Evidence from tennis matches*

3.1 Introduction

The best-known framework for rational decision making, expected utility theory, requires that probabilities and utilities are assigned to all relevant uncertain outcomes. Whilst a decision-maker aspiring to choose rationally might discover utilities by introspection, away from coin flips, dice rolls and spins of the roulette wheel, it is far from obvious which probabilities she should assign to uncertain outcomes. As Knight (2012) observed, the uncertainty attached to most interesting and important events is not readily quantifiable. Betting odds, turned into *implied probabilities*, and quoted by either bookmakers or by traders in prediction markets, offer the promise to turn many situations of Knightian uncertainty into situations of Knightian risk. This raises the question about the extent to which probabilities implied by betting odds possess the properties that make probabilities of risky events, such as those of coin flips, so reliable for decision making.

Betting odds are available for many events from a variety of domains and have been shown to predict outcomes well. Berg, Forsythe, Nelson, and Rietz (2008) investigate the accuracy of prediction markets for political outcomes, finding that they outperform polls, especially over longer horizons to elections. Cowgill and Zitzewitz (2015) study corporate prediction markets run by companies for their internal benefit and find they often outperform the predictions of executives. Dreber, Pfeiffer, Almenberg, Isaksson, Wilson, et al. (2015) use prediction markets to generate pre-

* I thank Thomas Dohmen, Hans-Martin von Gaudecker, Dominik Liebl, Thomas Neuber and Christian Zimpelmann for helpful comments and suggestions.

dictions that research in psychology will be reproducible, and similarly find that the resulting predictions perform well.

Whether prediction markets yield probabilities has also been investigated theoretically. Manski (2006) and Wolfers and Zitzewitz (2006) examined how the prices of predictions markets are related to the beliefs of traders in equilibrium. When traders have heterogeneous beliefs regarding the probabilities with which an event occurs, the equilibrium price of a prediction market equals the average of beliefs, provided traders are risk-averse with log utility. However, for risk-neutral traders and traders with different risk aversion parameters, this is no longer true, although Wolfers and Zitzewitz (2006) show prices will be close to mean beliefs under a range of plausible specifications. It is an open question whether mean beliefs have the desirable properties of probabilities of risky events, that is, whether they bear any relation to empirical events.

What makes the probability assessment $\frac{1}{2}$ reliable regarding the event "heads" when flipping a coin? This paper examines the following properties. First, over a large number of flips, we can be confident the fraction of heads will be approximately $\frac{1}{2}$; it is a well-calibrated prediction. Second, the probability assessment $\frac{1}{2}$ conditions on all relevant information, no additional information would alter our prediction. The remaining uncertainty is irreducible in non-experimental settings. We can summarise these properties as $\Pr(\text{Heads}|X) = \frac{1}{2}$ where X represents any information set. Using a sample of more than forty thousand tennis matches between 2002 and 2019, I systematically investigate the reliability of implied probabilities in the sense just described. I test whether $\Pr(Y|Q,\ X) = Q$ holds when Q represents the implied probability for a match outcome Y and X is some information set. When X is empty, this amounts to a calibration test. Calibration of betting odds has been studied before, Page and Clemen (2013) find that prediction market prices are overall quite well-calibrated in both sports and politics, but for events whose resolution is far in the future, they find evidence of substantial miscalibration. In particular, they find a favourite-longshot bias in which events unlikely according to implied probabilities occur less often empirically than they should, and likely events too often.

Betting odds calibration has also been examined for tennis matches, with Forrest and McHale (2007) finding evidence of a favourite-longshot bias. More recently, Lahvička (2014) studies bookmaker odds and Abinzano, Muga, and Santamaria (2016) odds from the Betfair exchange. Both corroborate and extend the earlier finding of a favourite-longshot bias. This paper extends the calibration analysis of Lahvička (2014) by using a more flexible regression model to detect miscalibration, and a calculation of implied probabilities from odds that is less susceptible to the favourite longshot bias.

In terms of analysing what information implied probabilities for tennis matches are conditional on, this paper extends the analysis of Kovalchik (2016) which is a systematic comparison of the predictive power of a variety of published prediction models as well as bookmaker odds for tennis matches in the year 2014. Kovalchik (2016) finds that bookmaker odds predict best, with model-based player ratings predicting next best. My analysis builds on Kovalchik (2016) in three ways. First, I consider match outcomes over a longer period, ranging from 2002 to 2019. Second, I use a much richer selection of player characteristics to serve as potential regressors in a prediction model. Third, I integrate bookmaker odds, player characteristics and model-based player ratings into an overall prediction using modern machine learning methods. In this sense, my analysis is similar to Groll, Ley, Schauberger, and Van Eetvelde (2019) who use the random forest model to combine characteristics of teams and models of team strength into an overall prediction in the context of predicting football matches, and to Gross and Rebeggiani (2018) who examine the efficiency of betting markets for football by contrasting odds with a statistical model. Overall, I find that implied probabilities for tennis matches are close to the ideal of Knightian risk, as exemplified by coin flip probabilities. First, they are almost perfectly calibrated in the sense that the empirical frequency of outcomes matches the implied probability. One notable exception is a favourite-longshot bias of a small magnitude for grand slam matches. Second, they achieve a lower prediction error than machine learning approaches based on a rich selection of publicly observable information. Third, models that use this information as well as implied probabilities as predictors do not predict better than implied probabilities alone. In terms of which machine learning methods perform best, I find that the more flexible regression models random forest and gradient boosting are outperformed by regularised logistic regression. This likely reflects that the strongest predictors of tennis matches, player ratings, are based on models of player strength that are by construction linear at the log odds scale.

In the next section, I describe the data sources. The empirical strategy is described in Section 3.3. Section 3.4 presents results and Section 3.5 concludes.

3.2 Data

To analyse the extent to which implied probabilities in tennis come close to probabilities of Knightian risk, I use two types of data: A database of tennis matches and players characteristics and betting odds offered by bookmakers.

3.2.1 Data sources

Public information with which to predict matches comes from an archive of results maintained by the ATP website[1] which has data going back to 1969. Since 1991, the information provided for each match has become substantially richer, going beyond match scores. For each match, the data additionally includes various indicators that capture players' performance: aces, double faults, first serve percentage, first serve

[1] Available from: https://www.atptour.com/en/scores/results-archive

points won, second serve points won, breakpoints saved. From these, detailed measures of each player's serving and returning performance can be calculated. I use a public data set collecting all information listed on the ATP website made available by Jeff Sackmann.[2]

For betting odds, I use data from the website Tennis-data.co.uk[3], which has been collecting a selection of the odds from different bookmakers shown on the website oddsportal.com since 2001. The number of matches for which data is available, and the years for which there are any data is available, differs across bookmakers. The longest series of odds is available for the bookmaker Bet 365, with thousands of matches between 2002 and 2019. These odds are the primary focus of the analysis.

3.2.2 Merging data sources

Match-identifying information differs for the two data sets. The match database contains the full names of the players, their ranks, the score of the match and the date of the start of the tournament, but not the dates of individual matches. The bookmaker data has exact dates for the matches, player ranks, match score, and last names, but only the first letter of first names. Across the data sets, there are some inconsistencies: Player names are sometimes differently written, match scores and ranks sometimes differ despite referring to the same match. Some of the identifying variables are missing in both data sets for some matches. To merge the data sets, I proceed in stages. First, I convert the full names from the match database to the same format as the names in the bookmaker data. Then matches are linked based on year, month, player names and score. Then I repeat this procedure with month lags, since the match database only contains the start of the tournament buts matched played in it could take in the following month. Finally, remaining matches are linked based on year, month and player ranks, also with month lags. After each stage of linking matches, I manually resolve duplicate connections. This procedure results in links for all but 236 bookmaker matches and 209 for the bookmaker Bet365, for which I carry out my analysis. For these matches, I identify as potential links those with the most similar[4] player names amongst the remaining unused matches in the database. I manually verify that links above a high name similarity threshold join data points corresponding to the same match. Following this final step, the number of bookmaker matches for which I do not have a link is 104, which is less than 0.2% of all bookmaker matches.

[2] Available here: https://github.com/JeffSackmann/tennis_atp

[3] The data was retrieved from: http://www.tennis-data.co.uk/alldata.php

[4] The similarity measure used is the product of similarity scores for each player returned by the SequenceMatcher algorithm implemented in Python's difflib module. To give an example, one match with names separated by 'v.' is 'wawrinka s. v navarro pastor i.' in the bookmaker data and 'wawrinka s. v navarro i.' in the match data, with a similarity of 0.74. The algorithm helps link matches when parts of the names are missing or spelt differently

Figure 3.1 shows the number of ATP Tour level tennis matches over time as well as those for which I have odds using this merging procedure. There are odds from bookmakers for a vast majority of ATP tour matches.

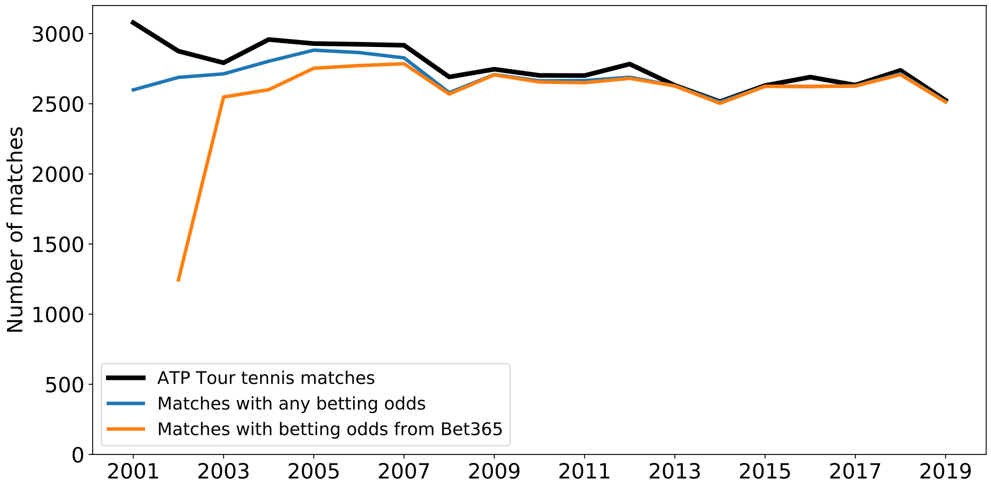


Figure 3.1. Number of ATP Tour tennis matches with and without betting odds over time

3.2.3 Calculation of implied probabilities

To define one outcome per match, I randomly and independently across matches assign one player to be player A and the other to be player B. I specify the outcome $Y = 1$ as the event that player A wins the match. Bookmakers offer customers bets on the outcomes of a match by quoting odds for each player. These odds specify the multiple of one's stake that is paid out if the player wins. If the player does not win, the stake is forfeit. To turn betting odds into a prediction for Y, that player A wins, I determine the subjective probability interval for which a risk-neutral gambler would not bet on either player and take the midpoint of this interval. The resulting implied probability q_A is always in the unit interval, and if the process were to be repeated for player B, $q_A + q_B = 1$. Thus implied probabilities calculated with this procedure satisfy the axioms of probability theory.

The details of the calculation are as follows. A bookmaker offers odds d_A for the event that player A wins and d_B for the event that player B wins. A risk-neutral agent with belief p_A that player A wins will bet on player A if the expected profit $d_A \cdot p_A - 1 \geq 0$ and on player B if $d_B \cdot (1 - p_A) - 1 \geq 0$. For beliefs p_A in the interval $\left[\frac{d_B - 1}{d_B}, \frac{1}{d_A}\right]$, a risk-neutral agent would thus not bet on either player. I define the implied probability q_A that player A wins as the midpoint of this interval.

Figure 3.2 gives an example of how betting odds for tennis matches were presented to prospective gamblers by the bookmaker Bet365 for the match Thiem v Zverev at the Australian Open in 2020.

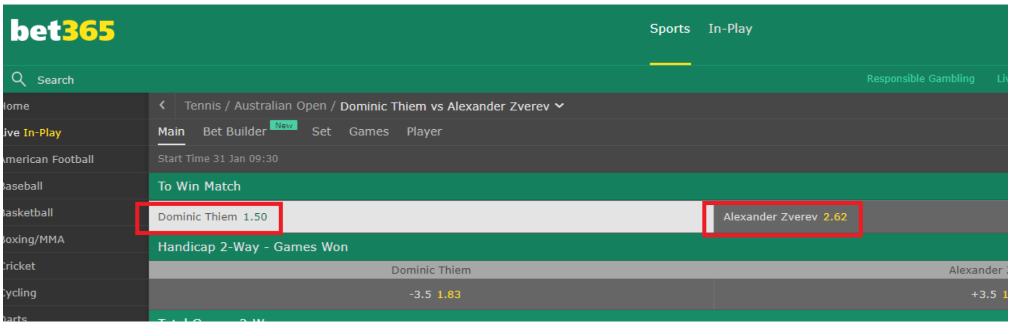


Figure 3.2. Example of betting odds from bookmaker Bet365

Notes: The odds refer to the semi final match Dominic Thiem (Austria) v Alexander Zverev (Germany) at the Australian Open 2020 on 31 January 2020. The numbers below 'To Win Match' are the decimal odds offered by bookmaker Bet 365 on Dominic Thiem (1.5) and Zverev (2.62). Since Thiem won, a bet of one unit on Thiem would have yielded a profit of 0.5 units. The implied probability of Thiem winning was 0.64.

Assigning Dominic Thiem to be player A, the odds offered for a bet on him were $d_A = 1.5$, the odds offered for a bet on Zverev were $d_B = 2.62$. A risk-neutral gambler would have refrained from placing a bet on either player if her subjective belief that Thiem wins had been located in $\left[\frac{2.62-1}{2.62}, \frac{1}{1.5}\right] = [0.62, 0.66]$. The midpoint of this interval, $q_A = 0.64$, is the implied probability.

In addition to yielding numbers that satisfy the probability axioms, this procedure has intuitive appeal through its connection to the subjective belief of a hypothetical trader who is indifferent at prevailing odds. It is however not the only procedure by which betting odds can be turned into values between $[0, 1]$ that sum to 1. Štrumbelj (2014) investigates the predictive performance of three different procedures: What they call the basic normalisation $q_A = \frac{\frac{1}{d_A}}{\frac{1}{d_A}+\frac{1}{d_B}}$, the predicted probability of a logistic regression of match outcomes on odds fit with historical data, and what they call Shin probabilities. Shin probabilities are based on a model due to Shin (1993) in which bookmakers set odds based on their subjective beliefs and the proportion of informed insider traders amongst gamblers. Using that model, odds offered by bookmakers can be inverted for the subjective probabilities they assign to the outcomes. Štrumbelj (2014) present evidence that Shin probabilities predict better than the two alternative procedures for outcomes across many sports. Despite the different motivations under which they are derived, for the case of two outcomes, the formula Štrumbelj (2014) gives for calculating Shin probabilities from odds produces values that are identical to the procedure I have described in this section.[5] I use them as implied probabilities throughout the rest of the paper. Lahvička (2014), who also examines calibration of implied probabilities based on the odds of Bet365, uses the basic normalisation.

3.3 Empirical strategy

In the first part of this section, I describe implications of the hypothesis that implied probabilities are probabilities of match outcomes that condition on a rich set of information and how these implications can be empirically tested. Key to the tests is a predictive model of tennis matches. In the second and third parts, I discuss the information sets and statistical models with which implied probabilities are compared.

3.3.1 Testing whether implied probabilities are probabilities of events and identifying the information they condition on

Define the outcome $Y = 1$ as the event that player A wins a tennis match. Let Q be the implied probability, calculated from bookmaker odds with the procedure described in the previous section. The assumption I want to test is that the implied probability is the probability that the event occurs, conditional on some information set Z. It can be stated as $Q = \Pr(Y = 1 \mid Z)$. If this assumption is true, it has several implications. Let X be any information set satisfying $Y \perp\!\!\!\perp X \mid Z$, that is X is independent from Y if we already know Z. Then $Pr(Y = 1 \mid Q, X) = Q$.[6] This holds for all values of $Q = q$ and $X = x$[7], and means that we can test whether implied probabilities are conditional probabilities with a regression of Y on the implied probability Q and observable information X. If implied probabilities are conditional probabilities, and the information they condition on is richer than X, then we expect to find that X does not change the probability of Y away from the implied probability, except due to estimation error. We can thereby indirectly draw the boundaries of the information set Z despite not observing it.

[5] Štrumbelj (2014) defines Shin probabilities for two outcomes in terms of the odds d_A and d_B as $q_{\text{A, Shin}} = \frac{\sqrt{z^2+4\cdot(1-z)\cdot\frac{\pi_A}{\beta}}-z}{2\cdot(1-z)}$ where $z = \frac{(\pi_+-1)(\pi_-^2-\pi_+)}{\pi_+(\pi_-^2-1)}$ and $\beta = \pi_+ = \pi_A + \pi_B$ and $\pi_- = \pi_A - \pi_B$ and $\pi_A = \frac{1}{d_A}$ and $\pi_B = \frac{1}{d_B}$. It is not readily apparent from inspection of these formulas, but numerically evaluating them for $d_A \geq 1$ and $d_B \geq 1$ shows they reduce to the midpoint of $\left[\frac{d_B-1}{d_B}, \frac{1}{d_A}\right]$

[6] This follows from the law of iterated expectations:

$$\begin{aligned} Pr(Y = 1 \mid Q, X) &= E[Y \mid Q, X] = E[E[Y \mid Z, Q, X] \mid Q, X]] = E[E[Y \mid Z] \mid Q, X]] \\ &= E[Q \mid Q, X] = Q \end{aligned}$$

[7] Representing the information set X with a vector of predictors

In theory, the sharpest test is obtained by making the information set X as rich as possible and searching for values $X = x$ at which $E[Y \mid Q = q, X = x] \neq q$. However, since $E[Y \mid Q, X]$ is unknown it has to be estimated, and its estimate can differ from q by chance or because the model underlying the estimate is misspecified. Nonparametric models that flexibly adapt to the data ward against misspecification, but when X is high dimensional, the number of observations in the neighbourhood of q, x will on average be small even in large samples and nonparametric estimation breaks down. As Athey and Imbens (2019) explain, a key lesson from the machine learning literature is that predictive power can often be increased by regularisation, which decreases estimation variance by introducing additional model inflexibility. This increases bias, but produces estimates of $E[Y \mid Q, X]$ that are on average closer to the true expectation due to the decrease in variance. Estimates can however be very biased for some values of x, q. Accordingly, it need not be true that $\widehat{E[Y|Q = q, X = x]} \neq q$ happens only by chance if the implied probability q fails to aggregate all the information contained in x. This might simply be one of the configurations of values x, q for which the estimator is very biased. When X is high dimensional, a sounder approach is thus to compare the overall predictive power $\widehat{E[Y|Q, X]}$ with that of Q using a measure of prediction error that is minimised by the true conditional probability.

Based on this insight, I implement the following steps to test whether implied probabilities are probabilities. First, I form $\widehat{E[Y \mid Q]}$ and use it to test $E[Y \mid Q = q] = q$ across the full range of values of q. Due to the large sample and low dimensionality, $\widehat{E[Y \mid Q]}$ can be modelled nonparametrically. The results of this analysis are in Section 3.4.1. Second, for a high-dimensional information set X, I form $\widehat{E[Y \mid X]}$ with machine learning model and compare its overall predictive power to that of Q. If $\widehat{E[Y|X]}$ predicts better than Q, then if $Q = E[Y \mid Z]$, we can conclude that X is richer than Z. The results of this analysis are in Section 3.4.2. Third, if $\widehat{E[Y|X]}$ predicts worse than Q, X might still contain elements not in Z, even though overall, it is not as rich and thus has inferior predictive power. To investigate this, I form $\widehat{E[Y \mid Q,X]}$ and compare its predictive power to Q. The results of this analysis are in Section 3.4.3.

These steps can be related to the concepts of calibration and sharpness Gneiting, Balabdaoui, and Raftery (2007) identify as key for evaluating probabilistic predictions. The first step is a calibration test, where calibration means the empirical frequency with which an event with probability assessment q occurs is q. Whilst a desirable property for a probability assessment to have, calibration on its own does not make for good predictions. The prediction 0.5, for instance, is always perfectly calibrated for the event that a randomly selected player will win a tennis match. Sharpness is the degree to which predictions differ from the uninformative value of 0.5. The second and third steps of the analysis use a large number of player features to compare the prediction error of implied probabilities and models. Prediction errors measured with proper scoring rules capture both calibration and sharpness (Gneiting, Balabdaoui, and Raftery (2007)).

3.3.2 Information set selection

The models of the second and third step of the analysis require a rich information set X. Constructing and selecting predictors that capture relevant information involves exploratory data analysis and particularly sophisticated predictors involve their own models. To ensure that statistical inference in the main analyses is unaffected, I used matches from the period 1991 to 2000 which precedes the data set of the main analysis. I group predictors into two broad categories, Player & Matchup Stats which are simple functions of player and match information in the database, and Ratings, which are model-based estimates of player strength.

Player & Matchup Stats
The data on players and historical matches published by the ATP offers a wide range of player attributes. For each player, the history of matches in which he was involved, with statistics on wins and losses, sets, points, and various serve and return related metrics are available. I refer to predictors capturing these variables, as well as statistics of the player matchup (previous results when player A encountered player B) as Player & Matchup Stats. When turning the raw data into predictors, I compute many variants of percentages such as the match win rate. The variants differ in the number of past matches over which they are calculated and the surface within which they pool matches (overall, hard, clay, grass or carpet). I also interact some of the variables with a dummy indicating whether the current match is a grand slam match. Grand slam matches are best-of-5 in the number of sets needed to win instead of best-of-3, which might make their outcomes more predictable conditional on the same information available about the players. A detailed explanation variables used and the variants calculated for the information set Player & Matchup Stats is given in Table 3.B.1.

Ratings
The analysis of Kovalchik (2016) shows that player ratings are the single best predictors of tennis matches. The statistics falling under Player & Matchup Stats, such as match win rates, do not account for the strength of players against which this was achieved. Since the strength of the opposition is also unknown, it has to be measured as well, which in turn depends on the strength of *their* opposition. Ratings are measures of a player's strength that systematically use information about the entire pool of players. Two simple measures of ratings I use are the sum of rank points awarded by the ATP for performing well in tournaments in a given season and the ATP rank, which is a 52-week moving sum of rank points. More sophisticated ratings

can be constructed based on latent variable models of match outcomes. The most widely used variant of such models is due to Elo (1978), bearing his name, which is used for the official rankings in chess and was the best performing forecast for tennis matches in the analysis of Kovalchik (2016). The ratings are based on the model $Y = 1\{r_A + u_A > r_B + u_B\}$ where r_A and r_B are latent ratings measuring player strength and u_A and u_B represent unmodeled components affecting whether player A wins. I calculate and use Elo and TrueSkill ratings.

Elo Rating

The Elo model assumes player ratings r_A, r_B are related to the probability of player A winning as $E[Y \mid r_A, r_B] = \frac{1}{1+10^{\frac{r_A - r_B}{400}}}$. This likelihood function amounts to assuming $u_B - u_A$ is logistically distributed. The ratings of each player are updated after match in which they are involved according to $r_{A,t+1} = r_{A,t} + \lambda \cdot (y - E[Y|r_{A,t} r_{B,t}])$. New players start with a rating of 1500[8]. The sign of the update depends on whether player A won and the magnitude on how surprising it was in light of the old rating as well as the parameter λ. This parameter determines rating volatility and thus how much weight the most recent match receives relative to previous matches. I implement the Elo model in Python and set λ to minimise the negative log-likelihood[9] when predicting match outcomes with the last available Elo rating prior to the match, using data between 1991 and 2000. I compute ratings for all matches, but optimise over matches from the last three years of this period in tournaments of the same level as in the match data for which I have betting odds. Ratings are not updated when a player retired with an injury because such results are not informative about player strength. In addition to ratings at the optimal λ, I compute ratings at different λ values around that value in case quickly changing (high λ) and slowly changing (low λ) ratings are in combination more predictive than any single value. I also calculate separate ratings for each playing surface (hard, clay, grass, carpet). Table 3.C.1 in the appendix contains the optimal λ values.

TrueSkill Rating

A more flexible alternative to the Elo model is TrueSkill, which has been developed by Herbrich, Minka, and Graepel (2007) for match-making in Microsoft's online games. TrueSkill ratings are a full Bayesian ratings, in the sense that a player's strength is modelled as a normally distributed random variable $r_A \sim \mathcal{N}(\mu_A, \sigma_A^2)$ which is updated by conditioning on match outcomes. TrueSkill ratings can be augmented with time dynamics that go beyond the Elo model, where these are implicitly controlled by λ. In TrueSkill r_A and r_B can be allowed to randomly drift over time. I model this as $r_{A,s} = r_{A,t} + (s-t) \cdot \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ where I set $s - t$

[8] The initial rating, as well as the $\frac{1}{400}$ constant are a convention. Though the optimal λ parameter is relative to the constant, the outcome probabilities would be identical for any choice as only score differences matter

[9] Defined as $-\sum_i y_i \cdot \log \hat{y}_i + (1-y_i) \cdot \log(1-\hat{y}_i)$ where $\hat{y}_i$ is the predicted probability that player A wins according to the model

to be the number of days between the beginning of the tournaments in which the two matches take place. From the perspective of the modeller, the rating drift leaves a player's expected rating unchanged but widens the dispersion around it. As a result, if two players face each other following a longer period of inactivity, it is reflected in the ratings that uncertainty over who will win increases. Player ratings are related to the probability of player A winning as $E[Y \mid r_A, r_B] = \Phi(\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B + \sigma_u})$ which corresponds to assuming that the unmodeled determinant of match outcomes satisfies $u \sim \mathcal{N}(0, \sigma_u^2)$. Ideally, one would update μ_A and σ_A according to Bayes' Rule following a math. Since no analytical solution is available for the update, I use the approximating equations for the two-player format from Dangauthier, Herbrich, Minka, and Graepel (2008). New players start with a prior rating $\mu = 0$ and $\sigma = 1$.[10] The parameters σ_u and σ_ϵ capture unpredictability of match outcomes conditional on ratings and how variable ratings are in periods of inactivity between time points respectively. I implement the TrueSkill model in Python and optimise both parameters for predictive power in the same fashion as for the Elo model. Section 3.C in the appendix gives the full details including the update equations. Table 3.C.1 contains the optimal parameter values.

I estimate regression models for three information sets: Player & Matchup Stats, Ratings and Full Info, which combines them and adds additional player variables such as age, height, playing hand and home advantage. For most variables, there is a measure both for player A, X_A and for player B, x_B. Because the identity of player A has been randomly assigned across all matches, they should affect the probability that player A wins with opposite signs but the same magnitude. By using only differences $x_A - x_B$, instead of separate measures, the dimensionality can be cut in half, at the cost of a functional form restriction. For the ATP world rank, for instance, $\text{logit}^{-1}(\beta(\log x_A - \log x_B))$ predicts much better than $\text{logit}^{-1}(\beta(x_A - x_B))$. For ATP ratings and ATP points, I log transform variables before computing the difference. For all other variables except dummies, I compute simple differences. Using match outcomes between 1991 and 2000, I also experimented with using x_A and x_B separately and found that it does not yield superior predictive power out of sample for any of the regression models. The number of predictors used in the Full Information specification is 474. Table 3.B.1 in the appendix summarises which predictors are included in each information set.

3.3.3 Model tuning

Since the objective is forming estimates of conditional probabilities, $\widehat{E[Y \mid Q, X]}$, I only consider models that minimise loss functions that are proper scoring rules

[10] As for the Elo model these are arbitrary, although the optimal parameters σ_u and σ_ϵ will change with the rating variance, as their magnitude is relative to it

because true probabilities uniquely minimise them. With binary outcomes, if the loss function L is a proper scoring rule, the prediction $h(x)$ based on predictors x that uniquely minimises expected loss $E[L(y, h(x))]$ is $h(x) = \Pr(Y \mid X = x)$ (Gneiting, Balabdaoui, and Raftery (2007)). I randomly partition my data set of tennis matches between 2002 and 2019, for which betting odds are available, into a training set and a test set. The training data set consists of 80% (36468) and the test data set of 20% (9112) of matches. When randomising, I stratify by year, tournament level and tournament round so that the test set has a balanced number of observations for these categories. This allows me to examine the heterogeneity of the main results.

Estimating $E[Y \mid Q]$

The first step of the analysis requires an estimate of $E[Y|Q]$. I estimate $\widehat{E[Y \mid Q = q]} = \text{logit}^{-1}(s(\text{logit}(q))$ where s is a penalised cubic spline, fit using the mgcv library in R (Wood (2011)). The smoothness of the spline, captured by the effective degrees of freedom, is automatically chosen based on the data using the restricted maximum likelihood criterion. Since Y is the outcome that player A wins, whose identity was randomly assigned across matches, $E[Y \mid Q] - Q$ must be symmetric around $Q = 0.5$. If it were the case that $E[Y \mid Q = 0.2] - 0.2 = -0.01$, that is, events with small implied probabilities occur somewhat less often than indicated by the implied probability, then it must also be the case that $E[Y \mid Q = 0.8] - 0.8 = 0.01$. If small implied probabilities are on average too large, then the complementary large implied probabilities must on average be too small. To obtain estimates that obey this symmetry, I take the midpoint of model predictions at the logit scale for q and the negative value of $1 - q$, prior to converting to the probability scale. This yields $\text{logit}^{-1}\left(\frac{s(\text{logit}(q)) - s(\text{logit}(1-q))}{2}\right)$ as the predicted probability of the outcome when the implied probability is q.

Estimating $E[Y \mid X]$ and $E[Y \mid Q, X]$

The second and third steps require estimates of the high dimensional conditional probabilities $E[Y \mid X]$ and $E[Y \mid Q, X]$, for which I estimate several machine learning models. Table 3.1 gives an overview of the models in terms of their functional forms and the hyperparameters that are optimised when fitting them.

The objective function optimised for all these models is the negative log-likelihood[11] which is a proper scoring rule, augmented with various penalties for model complexity. The first two models impose a linear and additive relation between log odds of outcomes and predictors, the third model an additive but component-wise possibly nonlinear relation modelled with splines, and the last two

[11] Mean negative log-likelihood $= -\frac{1}{n}\sum_i y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)$. This loss function is referred to with different names in the standard description of different models. For Random forests, it is referred to as entropy criterion, for gradient boosting models it is the binomial deviance

Table 3.1. Estimated models

Models	Functional form	Optimised hyperparameters
Ridge logistic regression	$\log \frac{p}{1-p} = \sum_k^K x_k \cdot \beta_j$	L2 penalty
Lasso logistic regression	$\log \frac{p}{1-p} = \sum_k^K x_k \cdot \beta_j$	L1 penalty
Generalised additive model	$\log \frac{p}{1-p} = \sum_k^K s_k(x_k)$	Combined L1 and L2 penalty
Gradient boosted classification trees	$p = \sum_j^J T(x; \theta_j) \cdot \beta_j$	Learning rate max tree depth, subsample ratio
Random forest classifier	$p = \frac{1}{B} \sum_b^B T(x; \theta_b)$	Min samples in tree leaf max tree depth, max features

Notes: p is the probability of the outcome, $s(\cdot)$ are splines, $T(\cdot)$ are decision trees. *L2* and *L1* penalise the magnitude of estimated coefficients with β^2 and $|\beta|$ respectively. In GAM, they penalise the coefficients of the basis expansion of predictors. Larger penalties decrease variance but increase bias. The tree ensembles gradient boosting and random forest have various hyperparameters, the listed parameters were selected following experimentation of what affects their predictive power. *Learning rate* scales the coefficient β_j with which a new tree T affects the overall prediction, with smaller values decreasing variance but increasing bias. *Max tree depth* controls how many times regressors x are split in a single tree before terminal nodes are reached. *Subsample ratio* controls the fraction of randomly selected observations used to fit an additional tree in each iteration. A smaller value decreases variance. *Min samples in tree leaf* determines the minimum observations necessary in both halves for tree splits. Larger values decrease variance but increase bias. *Max features* in Random Forest controls the number of randomly chosen predictors to be used for fitting each tree. A larger value makes the averaged trees more correlated but increases their predictive power. Full details of hyperparameter tuning are given in Section 3.D.

are ensembles of decision trees, which are sufficiently flexible to express complicated interactions and nonlinearities between the predictors and outcomes. Tree ensemble techniques tend do predict very well in a variety of settings[12], the random forest model, in particular, achieves a high degree of predictive power for football matches in the analysis of Groll, Ley, Schauberger, and Van Eetvelde (2019). These ensemble techniques use the base model of the decision tree $T(x, \theta)$ to form predictions. Decision trees partition the predictor space x by selecting the predictor and threshold such that splitting the sample in terms of that predictor and threshold, and making separate predictions in each half of the split, minimises a loss function. This is repeated with the resulting leaf nodes of the tree until a stopping criterion, defined by hyperparameters, is met. The parameters governing predictors and splits are collected in θ. Gradient boosting iteratively fits J trees such that the next tree is optimised to improve the overall prediction when combined with the predictions of the previously fit trees. In a random forest, each of B trees is fit using a new bootstrap sample and a randomly selected subset of predictors, and the final prediction averages across trees (Hastie, Tibshirani, and Friedman, 2009).

[12] https://www.import.io/post/how-to-win-a-kaggle-competition/

I use the training data set to select the best hyperparameters of the models with a grid search. All variables are standardised to mean zero and unit variance[13]. For the logit and GAM models, I use 10-fold cross-validation and choose the hyperparameters that minimise the cross-validation negative log-likelihood. For the gradient boosting model, I use 5-fold cross-validation and for random forests, I use error estimates based on out-of-bag predictions to economise on the computational burden. Since each tree model of a random forest is estimated with its own bootstrap sample of the data, some data points will not be included in it. Out-of-bag predictions for a data point average only trees not trained with this data point and are thus free from the downwards bias of in-sample error estimates (Hastie, Tibshirani, and Friedman, 2009). For the tree ensembles, I choose the hyperparameters listed in Table 3.1 and the range of values for the grid search following experimentation of what most affects most the predictive power of the models. After finding the best hyperparameters, I refit the models on the entire training set data. The GAM model is fit with the gamsel package in R (Chouldechova, Hastie, and Spinu, 2018)[14], gradient boosting trees are fit with the xgboost library (Chen and Guestrin, 2016) in Python and all other models are fit with the Scikit-learn library in Python (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, et al., 2011). More details and cross-validation results for the hyperparameter optimisation is given in Section 3.D. The prediction error estimates of Section 3.4.2 and Section 3.4.3 use the test data set only, which has not been used in any way to tune the models or specify the information sets. This ensures that they are unbiased.

3.4 Results

I this section, I present results for the three steps discussed in section Section 3.3. First, I test calibration, that is $E[Y \mid Q = q] = q$, second I compare the predictive power of $Q = E[Y \mid Z]$ and $E[Y \mid X]$, third, I compare the predictive power of $Q = E[Y \mid Z]$ and $E[Y \mid Q, X]$.

3.4.1 Testing $E[Y \mid Q = q] = q$ for all values q

I estimate $E[Y \mid Q]$ with the procedure described in the previous section for the implied probabilities of the bookmaker Bet365. Figure 3.3 shows the estimates $E[\widehat{Y \mid Q = q}] - q$ in aggregate, and separately for matches depending on the level of the tournament in which they took place, ranging from basic ATP 250 and ATP

[13] This ensures every predictor is on an even scale when its coefficient is penalised in the ridge, lasso and GAM models. It does not affect the tree ensemble models

[14] I also experimented with the GAM model of the mgcv package which I use to estimate $E[Y \mid Q]$, but found that it has worse predictive power.

500 tournaments to elite masters 1000 tournaments and grand slam tournaments (Australian Open, French Open, Wimbledon and US Open).

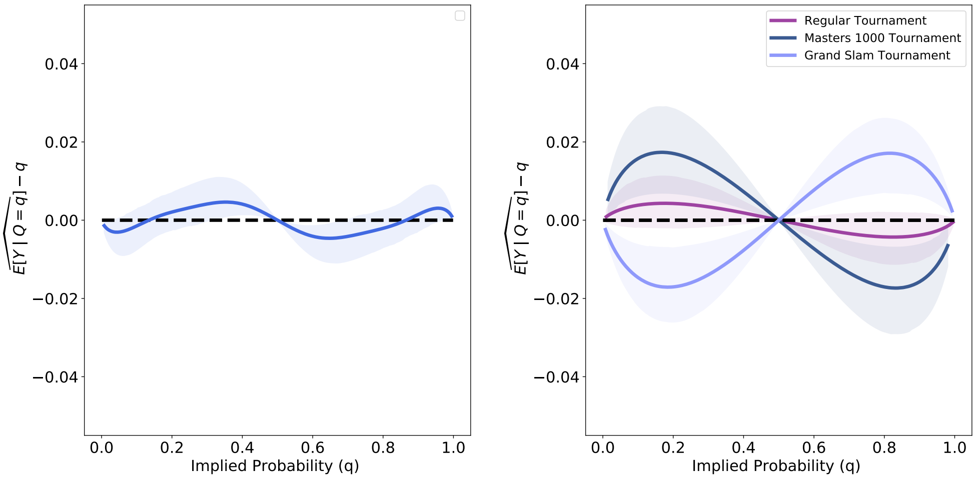


Figure 3.3. Testing $E[Y \mid Q] = Q$, implied probability from Bet365, aggregate and by tournament level

Notes: The lines are $\widehat{E[Y \mid Q = q]} - q$, the predicted probability that player A wins given implied probability Q for the values q of all matches. The left panel shows results pooled over all matches. In the right panel, separate lines are fit for levels of the tournament subject to the splines having the same effective degrees of freedom. Deviation of the predicted probability lines from the black dashed line measures the extent to which implied probabilities are miscalibrated for empirical frequencies. Estimates based on the model $\text{logit}^{-1}(s(\text{logit}(q))$ where s is a penalised cubic spline. To impose symmetry, the lines average the predictions for $s(\text{logit}(q))$ and $-s(\text{logit}(1-q))$ from this model before converting to the probability scale. Shaded regions are pointwise 95% confidence intervals obtained from 1000 bootstrap replications which include this averaging.

Noting the scaling of the y-axis which ranges from -5% to +5%, estimated deviations from perfect calibration are on average very small, and in line with what one would expect from chance. If the implied probability that player A wins is q per cent, player A goes on to win q per cent of the time across the full range of q.

When disaggregating matches by the level of tournament, there is some evidence of miscalibration of a small magnitude. For grand slam tournaments, the miscalibration is around 1.8 percentage points for longshots with implied probability at 0.2 and, by symmetry, for favourites at 0.8. The former win less often than the implied probability suggests, the latter more often. Interestingly, the opposite pattern occurs for matches from tournaments below those of grand slams in importance, but not for the lowest level tournaments. To check the robustness of these results, I fit the same regression for other bookmakers in my data set as well. Figure 3.A.2 in the appendix shows the results, reproducing the plot for Bet365 for comparison and ordering bookmakers by the number of observations available. The miscalibration for grand slam tournament occurs almost uniformly across bookmakers, despite the number of observations available, and years for which data is available, differing

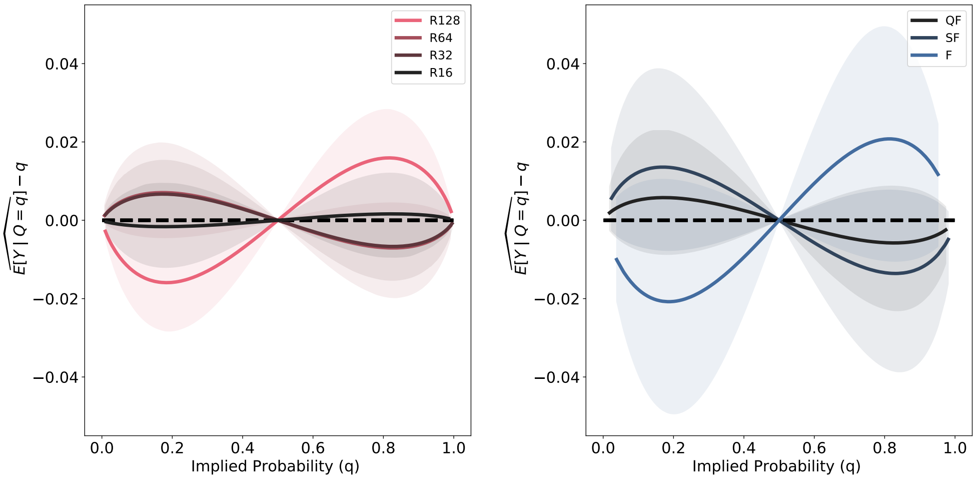


Figure 3.4. Testing $E[Y \mid Q] = Q$, implied probability from Bet365, by round of tournament

Notes: Separate lines are fit for each round of tournament subject to the constraint that the effective degrees of freedom are the same for each line. R128 refers to the first round of 128 players.

widely across them. The pattern for elite tournaments, by contrast, is less robust. The pattern for grand slam tournaments is an example of a well-known phenomenon in betting: The favourite-longshot bias. Ottaviani and Sørensen (2008) review possible explanations for this phenomenon. The explanations range from bettors misperceiving chances or having a preference for skewed returns (longshots), to bookmakers setting odds to protect themselves against informed gamblers as in the model of Shin (1993).[15] Figure 3.4 and Figure 3.5 disaggregates matches by the round of tournament and year respectively. There is no evidence that miscalibration has changed throughout the years, nor that it is more pronounced for different rounds of tournaments.

Using linear regression, Lahvička (2014) also examines implied probabilities of tennis matches from the bookmaker Bet365 for miscalibration, finding a favourite longshot bias overall, and a more pronounced version for grand slam tournaments and later rounds of tournaments. The different result compared to Figure 3.4 can be explained in terms of the difference in transformations to obtain implied probabilities from odds. Lahvička (2014) uses the basic normalisation (discussed in Section 3.2, see Štrumbelj (2014)) to transform odds into implied probabilities. Figure 3.A.1 in the appendix plots implied probabilities as computed in this paper against implied probabilities calculated with the basic normalisation, which shows that it produces consistently larger implied probabilities for longshots. Thus, if there

[15] Since the implied probabilities in my analysis coincide with what bookmakers believe if they set odds according to this model, it cannot, however, account for the present finding of a favourite-longshot bias.

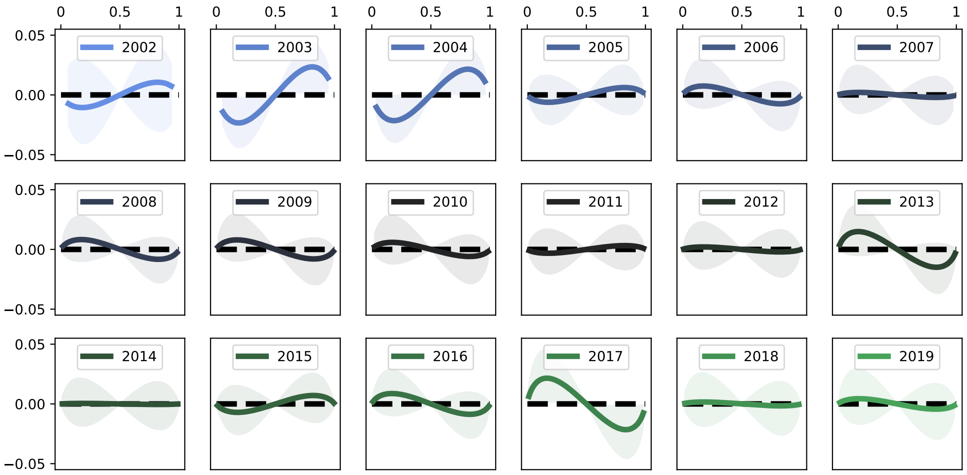


Figure 3.5. Testing $E[Y \mid Q] = Q$, implied probability from Bet365, by year

Notes: Separate lines are fit for each year subject to the constraint that the effective degrees of freedom are the same for each line.

was no miscalibration using the normalisation suggested by Štrumbelj (2014) and used in this paper, one could still find evidence of small miscalibration when using the basic normalisation.

3.4.2 Comparing prediction error of $Q = E[Y \mid Z]$ and $E[Y \mid X]$

I estimate $E[Y \mid X]$ for the models described in Section 3.3.3 for three information sets: Player & Matchup Stats contains variables that capture each player's performance or the history of the matchup, but do not adjust for strength of competition faced. Ratings are all variables involving the model-based Elo and TrueSkill measures as well as ATP points and the ATP rank. Finally, I use all predictors included characteristics such as age and player height, in the Full Info set. I use the implied probabilities from bookmaker Bet365 for calculating the predictor error of $Q = E[Y \mid Z]$. Table 3.2 shows the prediction errors, in terms of mean negative log-likelihood, of the optimised models for each information set. In Table 3.B.2 in the appendix, I show results for mean squared error, which are very similar. Error reductions of the best model over a 50% guess and of the implied probability prediction over the best model are also included.

Table 3.2. Prediction error by model and information set, compared to implied probability

Info set	Player & Matchup Stats	Ratings	Full Info
Model			
Ridge logistic regression	0.600	0.582	0.575
Lasso logistic regression	0.600	0.582	0.575
Generalised additive model	0.600	0.583	0.575
Gradient boosted cl. trees	0.601	0.584	0.578
Random forest classifier	0.604	0.586	0.581
Best model v guess	- 13.5%	- 16.0%	- 17.0%
Implied probability	0.568	0.568	0.568
Implied probability v best model	- 5.3%	- 2.5%	- 1.3% [0.000]

Notes: N = 9112 matches. Prediction error is measured with the mean negative log-likelihood, $-\frac{1}{n}\sum_i y_i \cdot \log \hat{y}_i + (1-y_i) \cdot \log(1-\hat{y}_i)$ where $\hat{y}_i$ is the predicted probability that player A wins from a model, or the implied probability calculated from betting odds. The means are computed over all matches in the test data set for which an implied probability from bookmaker Bet365 is available. Model were tuned with the training data by first using cross validation to find the best hyperparameters, then refitting the model with them. Player & Matchup Stats contains all variables capturing player statistics that do not adjust for strength of competition. Ratings contain ATP rank, ATP rank points, Elo and TrueSkill measures. Full Info uses all predictors. The last table cell has the p-value of a Diebold-Mariano test of equal predictive power in brackets.

The most important observation from Table 3.2 is that the best performing machine learning model that combines model-based player ratings and many other characteristics comes close, but ultimately still falls short of the implied probability. A Diebold-Mariano test of equal predictive power (Diebold and Mariano, 2002) indicates the difference is highly significant, even though it is small in magnitude. Implied probabilities from the bookmaker Bet365 lower the prediction error by 1.3% on the test data. What this suggests is that the unobservable information set Z that implied probabilities condition on is richer than the Full Info set of 474 predictors. The finding of Kovalchik (2016), that implied probabilities beat models, thus remains intact in my analysis. However, whereas the error reduction achieved by the best model (an Elo variant) over the implied probability in Kovalchik (2016) was 6.7%[16], the greater number of predictors combined with regularised regression models has considerably narrowed the gap.

Another observation is that the various models perform similarly but that tree ensembles do not outperform regularised logistic regression and the generalised additive model. The regularised logit models do best, with the type of penalty being irrelevant. The GAM model, which introduces additional flexibility for each predictor, allowing it to be nonlinear at the logit scale, achieves the same performance. The tree ensembles, which use decision trees as base models, do slightly worse. One potential explanation of this finding is that the most predictive features, ratings based on Elo and TrueSkill, are constructed to affect probabilities linearly in their difference at the log odds scale,[17] which is exactly how they enter the logit models. However, differences between the models are very small, with logit outperforming the random forest model by only about 1%.

Figure 3.6 subtracts from the prediction error of the best model (lasso logistic regression, but ridge logistic regression and GAM achieve effectively the same performance) for the full information set the prediction error of the implied probability. Separate results are given for each year, level of the tournament (regular, masters, grand slam) and round of the tournament (first round of 128 players (R128) to final).

The prediction error of the implied probability is lower than that of the best model throughout most years, and across all tournament levels and tournament rounds. There is a period between 2012 and 2014 in which the model does slightly better but this most likely due to chance, rather than due a genuine temporary decrease in predictive power of the implied probabilities. One particularly large gap in predictive error occurs at the first round of large tournaments. One potential explanation of this gap is that the most recent information used by the model to predict such matches will be based on how the players fared in the last tournament. The information aggregated by betting odds, Z, might additionally incorporate how the players have performed in training between tournaments.

[16] See Table 4 of Kovalchik (2016)

[17] For TrueSkill, it is the closely related probit scale

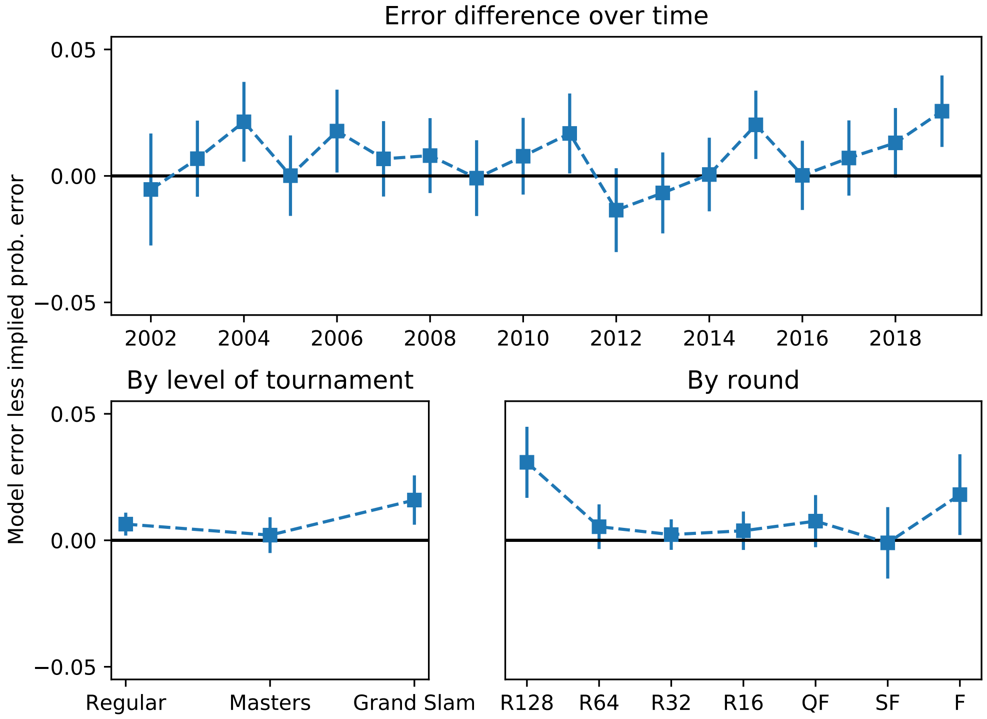


Figure 3.6. Heterogeneity in prediction error best model vs. implied probability

Notes: Plots show error differences between predictions from the best model (Lasso logistic regression) and implied probability (Bet365) over the test set matches. The categories of the top panel are years, those of the bottom left panel are levels of the tournament and those of the bottom right panel rounds. Prediction error differences are measured with the difference in mean negative log-likelihood $-\frac{1}{n}\sum_i y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)$ within each category. Vertical lines are pointwise 95% confidence intervals.

3.4.3 Comparing prediction error of $Q = E[Y \mid Z]$ and $E[Y \mid Q, X]$

The results of the previous section suggest that Z is richer than X, even when X is the Full Info set. This does, however, not mean that Z is necessarily a superset of X. Despite predicting better than X on average, there could be information in X that is not contained in Z. In this section, I re-estimate all models with Full Info plus the implied probability itself, that is, $E[Y \mid Q,X]$, to investigate this. Table 3.3 contains the results, Table 3.B.3 in the appendix shows the results for mean squared error. There is effectively no difference in the predictive power of Q and of $E[Y \mid Q, X]$ estimated with the best model. The null hypothesis of equal predictive power in the Diebold-Mariano test cannot be rejected (p=0.25). Ridge logistic regression, lasso logistic regression and the generalised additive model lower the prediction error compared to Q ever so slightly with the generalised additive model doing best. The tree ensembles, despite containing the implied probability Q as a predictor, do slightly worse than Q alone.

Table 3.3. Prediction error for model using implied probability as input

Info set Model	Full Info + Q
Ridge logistic regression	0.5679
Lasso logistic regression	0.5676
Generalised additive model	0.5672
Gradient boosted cl. trees	0.5691
Random forest classifier	0.5692
Implied probability	0.5680
Best model v Implied Probability	- 0.1% [0.25]

Notes: N = 9112 matches. Prediction error is measured with the mean negative log-likelihood $-\frac{1}{n}\sum_i y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)$ where $\hat{y}_i$ is the predicted probability that player A wins from a model, or the implied probability calculated from betting odds. Means are computed over all matches in the test data set for which an implied probability from bookmaker Bet365 is available. Model were tuned with the training data by first using cross validation to find the best hyperparameters, then refitting the model with them. Full Info + Q uses all predictors plus the implied probability. The last table cell has the p-value of a Diebold-Mariano test of equal predictive power in brackets.

Figure 3.7 shows heterogeneity in the prediction error of the best model less the prediction error of the implied probability across years, tournament levels and rounds. Their predictive performance is almost exactly the same across the categories; there is barely any evidence that X is not part of the information set Z that implied probabilities condition on. Though we cannot conclude that the uncertainty that remains conditional on Q is irreducible, because this analysis involved a number of modelling decisions that could potentially be improved, and does not use all information potentially available, we can conclude that it is very difficult to reduce it any further. The Full Info set containing 474 predictors does not reduce any uncertainty

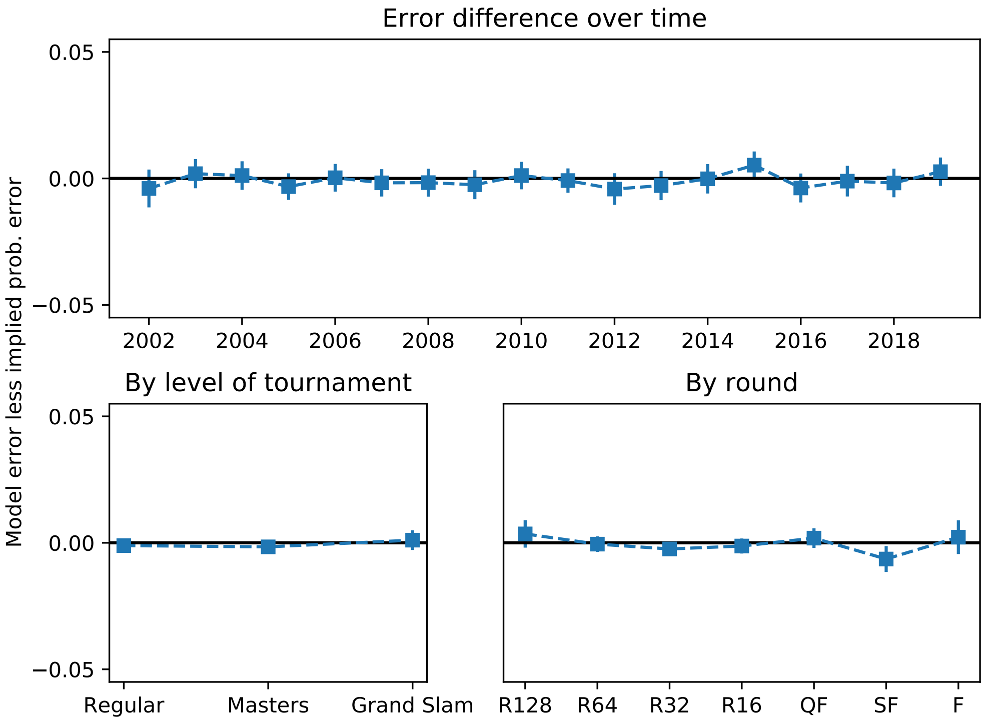


Figure 3.7. Heterogeneity in prediction error best model including implied probability as predictor v implied probability

Notes: Plots show error differences between predictions from the best model (generalised additive model) and implied probability (Bet365) over the test set matches. The categories of the top panel are years, those of the bottom left panel are levels of the tournament and those of the bottom right panel rounds. Prediction error differences are measured with the difference in mean negative log-likelihood $-\frac{1}{n}\sum_i y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)$ within each category. Vertical lines are pointwise 95% confidence intervals.

that remains when the implied probability is already in the model. In conjunction with the excellent calibration documented in Section 3.4.1, this suggests implied probabilities for tennis matches come close to the ideal of turning Knightian uncertainty into Knightian risk.

In terms of how the machine learning models differ when estimating $E[Y \mid Q, X]$, models that use the Lasso penalty (Lasso logistic regression, GAM) do best. If $Q = E[Y \mid Z]$ and $Y \perp\!\!\!\perp X \mid Z$, which is the hypothesis this section tests, then the conditional probability $E[Y \mid Q,X]$ is characterised by extreme sparsity, only a single predictor, Q, is not irrelevant. As Hastie, Tibshirani, and Friedman (2009) show, the Lasso penalty is particularly well suited for estimating such models because it can drive the contribution of irrelevant predictors down to exactly zero. Remarkably, the tree ensembles predict slightly worse than Q alone, despite containing Q as an input. This shows that model tuning with cross-validation does not fully prevent overfitting.

3.5 Conclusion

Using a combination of machine learning models and a rich selection of player features, I have examined whether implied probabilities have some of the desirable properties that probabilities in the domain of Knightian risk have: Whether they are calibrated, and whether they condition on an information set so rich the remaining uncertainty is irreducible.

I analyse calibration with nonparametric regression and find that on average, implied probabilities are very well calibrated across the full support of implied probabilities. Events with implied probability q per cent happen close to q per cent of the time, with deviations small enough to be consistent with being generated by chance. However, specifically examining grand slam matches, I corroborate earlier results and find that there is evidence of a favourite-longshot bias of a small magnitude. For those matches, players with implied probabilities in the region of 20% win only about 18% of the time. This pattern is robust across different bookmakers.

I analyse the information implied probabilities aggregate with hundreds of predictors that include player attributes, their track record in past matches and against the same opponent, and model-based player ratings. Fitting a variety of machine learning models, I estimate the probability of match outcomes with these predictors. I find that models come close to implied probabilities in predictive power, but still fall short. Implied probabilities lower the prediction error of the best model by 1.3% in terms of the negative log-likelihood. When adding implied probabilities as an additional predictor to the models, their predictive power is no greater than what is achieved by the implied probability alone. Together, these results suggest that implied probabilities are conditional on richer information than what is contained in the information set I have analysed. The uncertainty about outcomes that remains when knowing the implied probability cannot be reduced with it.

Comparing different machine learning models, I find that regularised logistic models perform slightly better than tree ensembles. This is likely the case because player ratings are the most predictive features in my analysis and they are constructed within a framework that makes them linear at the log odds scale. For regressions that use the full information set plus implied probabilities as predictors, I find that models using the Lasso (L1 penalty) fare best. This is likely because outcomes are independent from the entirety of the information once implied probabilities are conditioned on, which is a setting of extreme sparsity in which only a single predictor is relevant. Lasso regularisation excels in this setting because it can set the contribution of irrelevant predictors to exactly zero.

Overall, I find that in the case of tennis matches, implied probabilities come close to the ideal of turning Knightian uncertainty into Knightian risk.

Appendix 3.A Additional figures

Figure 3.A.1 compares the normalisation of betting odds used in the main analysis with an alternative.

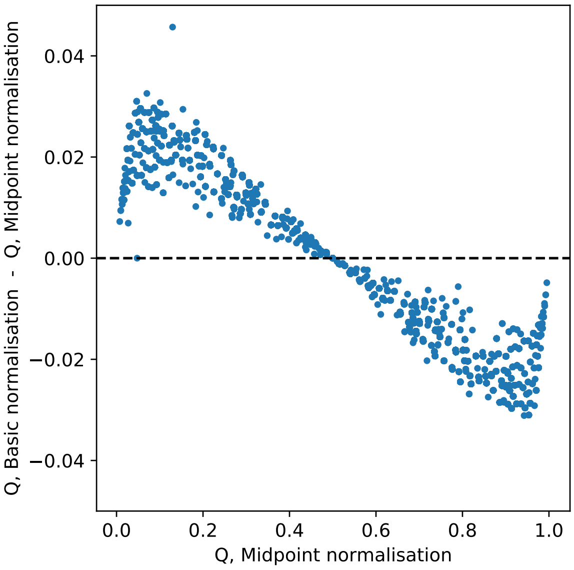
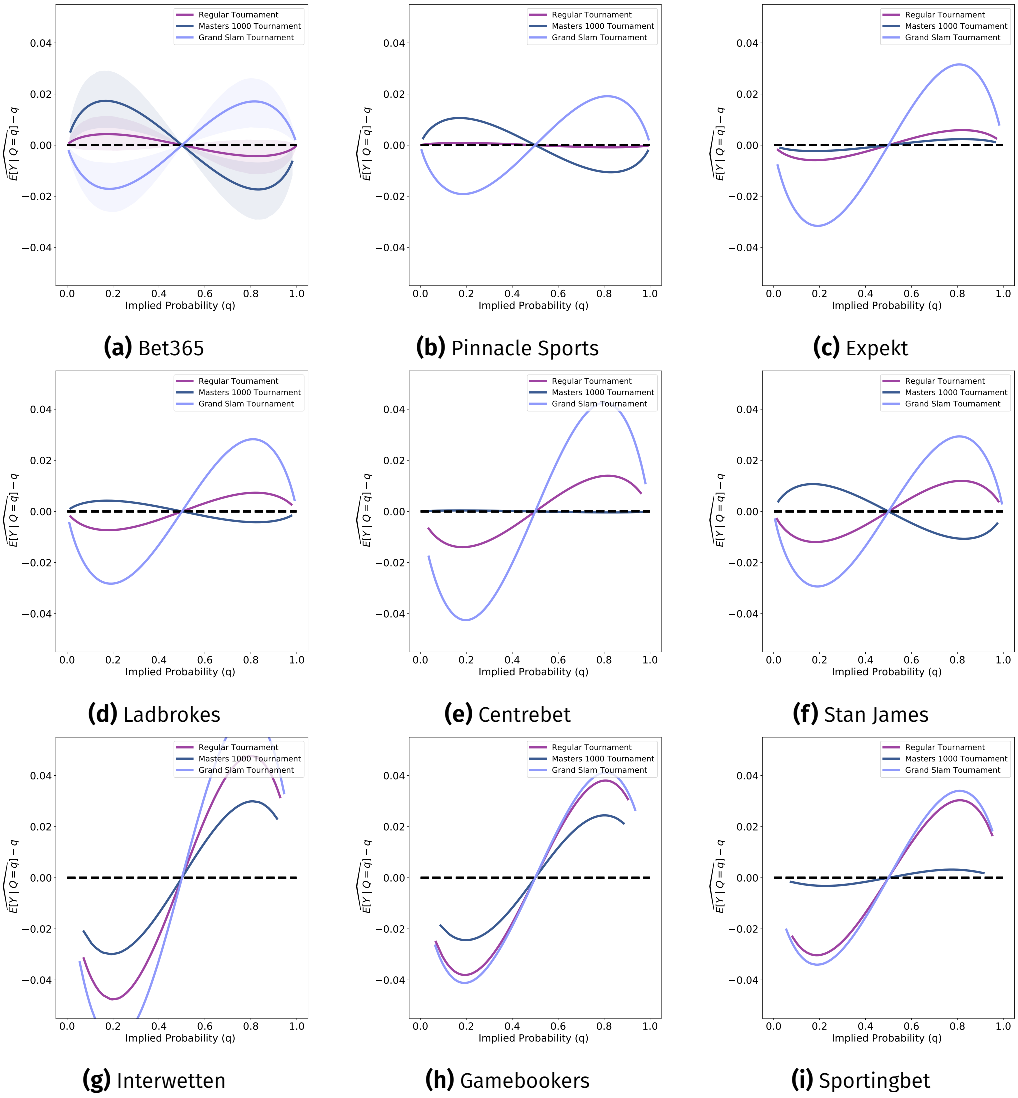


Figure 3.A.1. Implied probabilities: Midpoint normalisation vs. basic normalisation less midpoint normalisation

Notes: Midpoint normalisation is the procedure by which betting odds d_A, d_B are transformed into implied probabilities that is used in this paper, $q_A = \left[\frac{d_B - 1}{d_B}, \frac{1}{d_A}\right]$. Basic normalisation describes the alternative $q_A = \frac{\frac{1}{d_A}}{\frac{1}{d_A} + \frac{1}{d_B}}$. The scatter plot uses the midpoint normalisation on the x-axis, and subtracts it from the basic normalisation on the y-axis. For small implied probabilities on the midpoint normalisation, implied probabilities using the basic normalisation are between 2 and 3 percentage points larger. If midpoint based implied probabilities were perfectly calibrated, one would thus detect a favourite-longshot bias using the basic normalisation.

Figure 3.A.2 shows calibration analyses by level of the tournament for various bookmakers. Bookmakers differ in terms of the years for which observations are available and the total number of observations. The most robust form of miscalibration visible is a favourite longshot bias for grand slam tournaments.



(a) Bet365

(b) Pinnacle Sports

(c) Expekt

(d) Ladbrokes

(e) Centrebet

(f) Stan James

(g) Interwetten

(h) Gamebookers

(i) Sportingbet

Figure 3.A.2. Testing $E[Y \mid Q] = Q$ by tournament level for different bookmakers

Appendix 3.B Additional tables

Table 3.B.1. Variables used in information sets

Included in… Base variable	Player & Matchup Stats	Ratings	Full Info
Age	No	No	Yes
Years played	No	No	Yes
Right handed	No	No	Yes
Home advantage	No	No	Yes
Height	No	No	Yes
Player Rank	No	Yes	Yes
Player Ranking Points	No	Yes	Yes
Elo Rating	No	Yes	Yes
TrueSkill Rating	No	Yes	Yes
Match win percentage	Yes	No	Yes
Set win percentage	Yes	No	Yes
Point win percentage	Yes	No	Yes
Service points win percentage	Yes	No	Yes
First serve win percentage	Yes	No	Yes
Second serve win percentage	Yes	No	Yes
Break points saved percentage	Yes	No	Yes
Service game win percentage	Yes	No	Yes
Aces per match	Yes	No	Yes
Serve Rating	Yes	No	Yes
Double faults per match	Yes	No	Yes
Return points win percentage	Yes	No	Yes
First serve return win percentage	Yes	No	Yes
Second serve return win percentage	Yes	No	Yes
Return games win percentage	Yes	No	Yes
Break point win percentage	Yes	No	Yes
Return Rating	Yes	No	Yes
Matches played in current season	Yes	No	Yes
Hours played in current season	Yes	No	Yes
Sets played in current season	Yes	No	Yes
Points played in current season	Yes	No	Yes
Hours played in current tournament	Yes	No	Yes
Sets played in current tournament	Yes	No	Yes
Points played in current tournament	Yes	No	Yes
Days since last match	Yes	No	Yes

Notes: Information sets contain features constructed from the player specific base variables listed in the table. *Player & Matchup Stats* calculates expanding rates and rolling rates with windows of match length 80, 40, 20, 10, 3, 2, and 1. Rates are calculated pooling across matches and by surface. Match, set and point win percentages are also calculated with expanding match windows within tournaments. Match and set win percentages are also calculated within player matchups for expanding and 5 match windows. The number of predictors is 351. *Ratings* includes Elo ratings variants with update parameter λ values: optimal, 5, 10, 15, 30, 35, 40, 50, 60, 70, 80, 90, pooling across matches and within surface. Both ratings and outcome probabilities calculated from the Elo model likelihood are included. TrueSkill ratings include the mean and standard deviation of players strength and the outcome probability calculated from the model likelihood. The number of features is 116. *Full info* combines the other info sets and adds basic player attributes such as age. The number of features is 474. Ratings variables, as well as win, point and set percentages, are interacted with a dummy indicating whether the current match is in a grand slam tournament.

Table 3.B.2 and Table 3.B.3 replace the mean negative log-likelihood with the mean squared error when comparing the predictive power of models and implied probabilities.

Table 3.B.2. Prediction error by model and information set, mean squared error

Info set Model	Player & Matchup Stats	Ratings	Full Info
Ridge logistic regression	0.207	0.200	0.197
Lasso logistic regression	0.207	0.200	0.197
Generalised additive model	0.207	0.200	0.197
Gradient boosted cl. trees	0.208	0.201	0.198
Random forest classifier	0.209	0.201	0.199
Best model v guess	- 17.3%	- 20.1%	- 21.4%
Implied probability	0.194	0.194	0.194
Implied probability v best model	- 6.3%	- 3.1%	- 1.5% [0.000]

Notes: N = 9112 matches. Prediction error is measured with the mean squared error $\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$ where $\hat{y}_i$ where $\hat{y}_i$ is the predicted probability that player A wins from a model, or the implied probability calculated from betting odds. The means are computed over all matches in the test data set for which an implied probability from bookmaker Bet365 is available. Model were tuned with the training data by first using cross validation to find the best hyperparameters, then refitting the model with them. Player & Matchup Stats contains all variables capturing player statistics that do not adjust for strength of competition. Ratings contain ATP rank, ATP rank points, Elo and TrueSkill measures. Full Info uses all predictors. The last table cell has the p-value of a Diebold-Mariano test of equal predictive power in brackets.

Table 3.B.3. Prediction error for models using implied probability as input, mean squared error

Info set Model	Full Info + Q
Ridge logistic regression	0.1936
Lasso logistic regression	0.1934
Generalised additive model	0.1932
Gradient boosted cl. trees	0.1939
Random forest classifier	0.1940
Implied probability	0.1935
Best model v Implied Probability	- 0.2% [0.27]

Notes: N = 9112 matches. Prediction error is measured with the mean squared error $\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$ where $\hat{y}_i$ is the predicted probability that player A wins from a model, or the implied probability calculated from betting odds. Means are computed over all matches in the test data set for which an implied probability from bookmaker Bet365 is available. Model were tuned with the training data by first using cross validation to find the best hyperparameters, then refitting the model with them. Full Info + Q uses all predictors plus the implied probability. The last table cell has the p-value of a Diebold-Mariano test of equal predictive power in brackets.

Appendix 3.C Additional details for rating systems

Table 3.C.1 shows the optimal parameter values for the rating systems Elo and TrueSkill.

Table 3.C.1. Optimised parameters of ratings

Rating	Elo	TrueSkill	
Parameter	λ	σ_u	σ_ε
Surface			
Pooled	48.4	3.137	0.089
Hard	54.4	3.153	0.073
Clay	55.4	3.141	0.091
Grass	41.9	3.168	0.032
Carpet	41.9	3.168	0.050

Notes: Table shows the optimised parameter values for the Elo and TrueSkill ratings pooling all matches and by surface. λ determines the magnitude of the rating update for Elo. σ_u is the standard deviation of the random component determining matches in TrueSkill. σ_ε is the standard deviation of the mean-zero error term that is added to player ratings for every day between the starts of tournaments.

The TrueSkill model initialises player A and B with ratings $r_A \sim \mathcal{N}(\mu_{A,t}, \sigma^2_{A,t})$ and $r_B \sim \mathcal{N}(\mu_{B,t}, \sigma^2_{B,t})$. The outcomes of their match is modelled as $Y = 1\{r_A + u_A > r_B + u_B\}$ where $u_B - u_A \sim \mathcal{N}(0, \sigma^2_u)$. Upon observing the realised value of Y, the distribution of the player ratings needs to be updated. Herbrich, Minka, and Graepel (2007) develop approximate solutions for the update. The case for two players with the possibility of a draw is given in Dangauthier, Herbrich, Minka, and Graepel (2008) and reproduced here, setting the probability of draws to zero. Suppose $Y = 1$, that is, player A won, then updates are:

$$\mu_{A,t+1} = \mu_{A,t} + \frac{\sigma^2_{A,t}}{c} \cdot v\left(\frac{\mu_{A,t} - \mu_{B,t}}{c}\right)$$

$$\sigma_{A,t+1} = \sigma_{A,t} \cdot \sqrt{1 - \frac{\sigma^2_{A,t}}{c^2} \cdot w\left(\frac{\mu_{A,t} - \mu_{B,t}}{c}\right)}$$

Where $c^2 = \sigma^2_u + \sigma^2_{A,t} + \sigma^2_{B,t}$, $v(x) := \frac{\Phi(x)}{\phi(x)}$ and $w(x) := v(x)\dot{(}v(x) + x)$, with Φ the cdf andϕ the pdf of the standard normal distribution. The update for player B reverses the sign of the change for the mean parameter.

Between matches, I increase the standard deviation of player ratings by

$$\sigma_{A,s} = \sqrt{\sigma^2_{A,t} + (s-t)\sigma^2_\epsilon}$$

where $s - t$ is the number of days between the starts of tournaments in which the matches take place. For matches within tournaments, the only change to the standard deviation of player ratings is due to updating based on match results. μ and

σ of the player ratings are initialised at 0 and 1, the parameters σ_u^2 and σ_ϵ^2 are chosen to minimise the negative log likelihood when using ratings to predict match outcomes.

Appendix 3.D Hyperparameter tuning

For all models, hyperparameters are chosen by finding the values of the parameters that minimise an estimate prediction error (mean negative log likelihood) that does not automatically decrease with the flexibility of the model. For ridge and lasso regression there is a single hyperparameter c, which governs the importance of the log likelihood relative to the penalty of the coefficients, with larger values of c corresponding to less regularisation. For the GAM model, I optimise the hyperparameter λ which governs the importance of the penalty component of the objective function with larger values corresponding to more regularisation. The estimate of the prediction error is based on the hold-out folds in 10-fold cross-validation.

For gradient boosting, the optimised hyperparameters are the learning rate (10 values, evenly spaced between 0.02 and 0.2), the maximum depth of trees (2 or 3) and the subsample fraction (0.8 or 1). All other hyperparameters are left at the default values of the xgboost library. The parameters selected for optimisation and their range was set following experimentation of what affects the predictive power of the model. The estimate of the prediction error is based on the hold-out folds in 5-fold cross-validation. Trees are added until an early stopping condition is met: Cross-validation error fails to decrease at least every 10th new tree.

For random forests, the optimised hyperparameters are the maximum depth of trees (5, 10, None), the maximum number of randomly selected predictors for each tree ($\sqrt{K}$, $0.2\dot{K}$, $0.5\dot{K}$, None. K is the number of predictors, None means all K predictors are used.) and the minimum samples in tree leaves (1, 50, 100). All other hyperparameters were left at the default values of the scikit-library. The parameters selected for optimisation and their range was set following experimentation of what affects the predictive power of the model. The estimate of the prediction error is based on out-of-bag predictions. Each tree $T(x, \theta)$ in a random forest is fit using a bootstrap sample of the data. The probability that a data point is used by a tree is thus $1 - (1 - \frac{1}{n})^n$, which in a large sample is approximately $1 - \exp(-1) \approx 0.63$. For each data point, the out-of-bag prediction of a random forest averages the predictions of those trees that were not fit with it. This estimator of prediction error is comparable to cross-validation in that it does not automatically decrease with model flexibility, but requires less computational resources to implement (Hastie, Tibshirani, and Friedman, 2009).

Figure 3.D.1 and Figure 3.D.2 show the error estimates of hyperparameter values for the ridge and lasso logistic regression models and the generalised additive model respectively. Table 3.D.1 and Table 3.D.2 list the best hyperparameter combinations and error estimates for gradient boosting and random forests respectively.

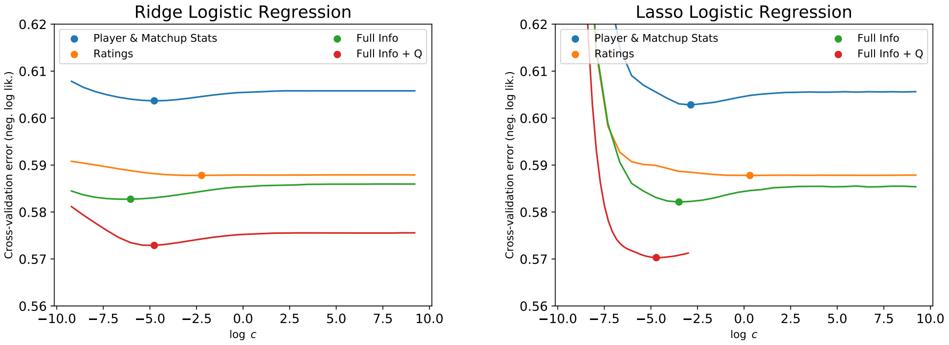


Figure 3.D.1. Cross-validation errors for hyperparameters of ridge and lasso logistic regression

Notes: Each curve shows the cross-validation mean negative log likelihood for an information set for different values of the hyperparameter. Log scaling of the x-axis showing values of the hyperparameter c. Higher values of c decrease regularisation. The Full Information + Q curve stops before large values of c are reached for the Lasso logit model because I set a grid of values over a smaller range. This was done because the inclusion of the implied probability Q makes it likely that more $L1$ regularisation would be optimal, which turns out to be the case. The marker of each line identifies the best value of the hyperparameter.

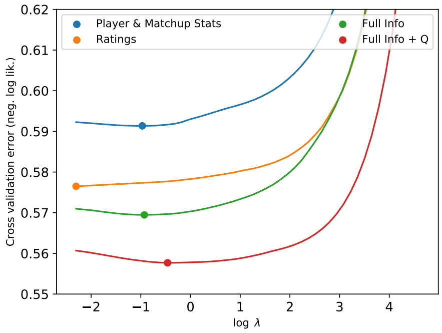


Figure 3.D.2. Cross-validation errors for hyperparameters of generalised additive model

Notes: Each curve shows the cross-validated mean negative log likelihood for an information set for different values of the hyperparameter. Log scaling of the x-axis which shows values of the hyperparameter λ. Higher values increase regularisation. The marker of each line identifies the best value of the hyperparameter.

Table 3.D.1. Cross-validation error of best hyperparameters of gradient boosted classification trees

Info set Hyperparameter	Player & Matchup Stats	Ratings	Full Info	Full Info + Q
Learning rate	0.04	0.06	0.04	0.08
Max depth	2	3	2	2
Subsample ratio	0.8	0.8	0.8	0.8
Error	0.607	0.590	0.585	0.571

Notes: Table shows the best hyperparameter combinations for each information set according to cross-validated mean negative log likelihood for each information set.

Table 3.D.2. Out-of-bag error for best hyperparameters of random forest

Info set Hyperparameter	Player & Matchup Stats	Ratings	Full Info	Full Info + Q
Max depth	10	10	None	5
Max features	0.5	None	None	None
Min samples leaf	50	100	100	1
Error	0.611	0.591	0.588	0.573

Notes: Table shows the best hyperparameter combinations for each information set according to out-of-bag mean negative log likelihood for each information set.

References

Abinzano, Isabel, Luis Muga, and Rafael Santamaria. 2016. “Game, set and match: the favourite-long shot bias in tennis betting exchanges.” *Applied Economics Letters* 23 (8): 605–8. [142]

Athey, Susan, and Guido W Imbens. 2019. “Machine learning methods that economists should know about.” *Annual Review of Economics* 11: 685–725. [148]

Berg, Joyce, Robert Forsythe, Forrest Nelson, and Thomas Rietz. 2008. “Results from a dozen years of election futures markets research.” *Handbook of Experimental Economics Results* 1: 742–51. [141]

Chen, Tianqi, and Carlos Guestrin. 2016. “Xgboost: A scalable tree boosting system.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. [154]

Chouldechova, Alexandra, Trevor Hastie, and Vitalie Spinu. 2018. “gamsel: Fit regularization path for generalized additive models.” *R package version* 1 (1): [154]

Cowgill, Bo, and Eric Zitzewitz. 2015. “Corporate prediction markets: Evidence from google, ford, and firm x.” *Review of Economic Studies* 82 (4): 1309–41. [141]

Dangauthier, Pierre, Ralf Herbrich, Tom Minka, and Thore Graepel. 2008. “Trueskill through time: Revisiting the history of chess.” In *Advances in Neural Information Processing Systems*, 337–44. [151, 168]

Diebold, Francis X, and Robert S Mariano. 2002. “Comparing predictive accuracy.” *Journal of Business & Economic Statistics* 20 (1): 134–44. [158]

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A Nosek, and Magnus Johannesson. 2015. “Using prediction markets to estimate the reproducibility of scientific research.” *Proceedings of the National Academy of Sciences* 112 (50): 15343–47. [141]

Elo, Arpad E. 1978. *The rating of chessplayers, past and present.* Arco Pub. [150]

Forrest, David, and Ian McHale. 2007. “Anyone for tennis (betting)?” *European Journal of Finance* 13 (8): 751–68. [142]

Gneiting, Tilmann, Fadoua Balabdaoui, and Adrian E Raftery. 2007. “Probabilistic forecasts, calibration and sharpness.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 (2): 243–68. [148, 149, 152]

Groll, Andreas, Cristophe Ley, Gunther Schauberger, and Hans Van Eetvelde. 2019. “A hybrid random forest to predict soccer matches in international tournaments.” *Journal of Quantitative Analysis in Sports* 15 (4): 271–87. [143, 153]

Gross, Johannes, and Luca Rebeggiani. 2018. “Chance or Ability? The Efficiency of the Football Betting Market Revisited.” Working Paper. [143]

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media. [153, 154, 162, 170]

Herbrich, Ralf, Tom Minka, and Thore Graepel. 2007. “TrueSkill™: a Bayesian skill rating system.” In *Advances in Neural Information Processing Systems*, 569–76. [150, 168]

Knight, Frank H. 2012. *Risk, uncertainty and profit*. Courier Corporation. [141]

Kovalchik, Stephanie Ann. 2016. “Searching for the GOAT of tennis win prediction.” *Journal of Quantitative Analysis in Sports* 12 (3): 127–38. [142, 143, 149, 150, 158, 159]

Lahvička, Jiri. 2014. “What causes the favourite-longshot bias? Further evidence from tennis.” *Applied Economics Letters* 21 (2): 90–92. [142, 147, 156]

Manski, Charles F. 2006. “Interpreting the predictions of prediction markets.” *economics letters* 91 (3): 425–29. [142]

Ottaviani, Marco, and Peter Norman Sørensen. 2008. “The favorite-longshot bias: An overview of the main explanations.” In *Handbook of Sports and Lottery markets*. Elsevier, 83–101. [156]

Page, Lionel, and Robert T Clemen. 2013. “Do prediction markets produce well-calibrated probability forecasts?” *Economic Journal* 123 (568): 491–513. [142]

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. “Scikit-learn: Machine learning in Python.” *Journal of Machine Learning Research* 12 (Oct): 2825–30. [154]

Shin, Hyun Song. 1993. “Measuring the incidence of insider trading in a market for state-contingent claims.” *Economic Journal* 103 (420): 1141–53. [146, 156]

Štrumbelj, Erik. 2014. “On determining probability forecasts from betting odds.” *International Journal of Forecasting* 30 (4): 934–43. [146, 147, 156, 157]

Wolfers, Justin, and Eric Zitzewitz. 2006. “Interpreting prediction market prices as probabilities.” Working paper. National Bureau of Economic Research. [142]

Wood, Simon N. 2011. “Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (1): 3–36. [152]