Essays in Behavioral Economics and Econometrics

Dissertation

zur Erlangung des akademischen Grades
doctor rerum politicarum
(Doktor der Wirtschaftswissenschaft)

eingereicht an der

Wirtschaftswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

Dipl.-Volksw. Christian Zankiewicz

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Wirtschaftswissenschaftlichen Fakultät:
Prof. Dr. Christian D. Schade

Gutachter:

1. Prof. Georg Weizsäcker, Ph.D
2. PD Dr. Yves Breitmoser

Tag des Kolloquiums : 28. August 2017

Acknowledgments

I am extremely grateful to my supervisor and co-author Georg Weizsäcker for his support during my PhD studies. Throughout the years he provided me with a lot of encouragement, many helpful suggestions for improvement, and advice when I needed it. I have learned a lot from him.
Yves Breitmoser, too, provided a lot of support and very helpful comments at various stages during the PhD. I am grateful to call him my second supervisor. I was also lucky to work with Ludwig Ensthaler and Tobias Schmidt on exciting projects. An additional source of inspiration were the people of the DIW Graduate Center Berlin. Of the 14 total strangers, who started their PhD studies together with me, I call all of them my friends today. I also quickly became fond of the girl with whom I had the privilege to share an office. It was anything but boring with her.
Another great thanks goes to the entire Competition and Consumers Department at DIW Berlin and especially to its former head Pio Baake. Furthermore, I acknowledge financial support from the European Research Council (Starting Grant 263412).

At last, I would like to thank those people who might not ever even read this but who made me who I am, my family.

Summary

The core of most standard economic theories about human behavior is based on a simple model of rational decision making. Rational individuals maximize a utility function by correctly processing all available information. They behave time-consistent, self-interested, and are not influenced by the choice environment.

Evidence from behavioral and experimental economics, however, suggests that actual human behavior exhibits many violations of this concept. People do not seem to be like those perfectly rational economic agents living in economics text books. If anything, they behave time-inconsistent (Thaler, 1981), they care about the welfare of others (Charness and Rabin, 2002), they are influenced by their temporary emotions (Loewenstein and Lerner, 2003), and they resort to heuristics when faced with a complex choice environment (Gabaix et al., 2006), among others.

In the line with this literature, the three chapters of this dissertation shed light on different aspects of human behavior that are at odds with rational behavior. Each chapter contributes to the existing behavioral economic research using either experimental, empirical, or methodological tools. First, by proposing and experimentally testing a simple behavioral model that extends the literature on the misperception of multiplicative growth processes, Chapter 1 of this dissertation aims to explain common money mistakes that people often make with long-term investments such as retirement savings plans. Second, in Chapter 2, real-life investment data of an online-lending platform are used to empirically investigate if private investors behave as the standard economic literature would predict and solely consider an investment's expected return or if they also care about other non-financial attributes of a debtor. The focus of the analysis is on gender discrimination, thereby defining and econometrically testing different concepts of how investors discriminate between male and female borrowers. Third, Chapter 3 takes a methodological path and proposes a novel experimental design that accounts for the empirically well-documented difficulties that survey respondents typically have when asked to state subjective probabilities. A binary choice approach embedded in an adaptive experimental design helps to minimize effort of the respondents, thus allowing for a more practical belief elicitation in both the lab and the field.

Chapter 1. In the first chapter, one of the most widely cited biases in the behavioral finance literature on the perception of multiplicative growth processes, usually referred to as "exponential growth bias," is extended from the deterministic to the stochastic domain. In its original deterministic version, it describes the failure to compound the effects of multi-period growth, for a given growth rate. For example, when Stango and Zinman (2009) ask study participants to forecast the total effect of accumulating 7% growth for ten periods, a substantial fraction of respondents gives an answer that is closer to 70% than to the actual 97%. But this is just half of the story, when the exponential growth bias is extended to stochastic settings it can be split up in the above described tendency to linearize (labeled "linearity bias" hereafter) and a second bias that will be referred to as "skewness neglect," describing an ignorance for the skewness of the outcome distribution that arises from the compounding of random growth over time.

A first series of incentivized laboratory experiments examines both of the above-described biases systematically and finds that the participants' perception of stochastic growth deviates in predictable ways from the rational prediction. Both of the above biases are found to be relevant. Overall, the experimental results are in line with a simple model of misperceiving compound shocks. This model, which is labeled as the "exponential growth bias model," has the agent perceive growth as a linear process, in the sense that all multiplicative growth is mistaken as additive growth. The model predicts linearity bias and skewness neglect.

A second series of experiments tests the predictions of the model in different variations of the experimental setting. It varies the incentive schemes, the level of feedback, as well as the nature of both the investment strategy and the underlying asset that the participants are asked to assess. Thereby, the robustness of the effects can be investigated as well as several other implications

of the exponential growth bias model. Qualitatively, almost all predictions of the model are borne out in the data, and often with large discrepancies from the rational prediction. This holds both for experimental treatments concerning more abstract stylized growth processes and for treatments employing a class of more realistic assets that are based on the historical returns of the German DAX index. In the latter set of experiments, assets are designed to emulate leveraged ETFs where the DAX index is the underlying asset. Such leveraged products were popular with household investors in the U.S. in recent years (though with a different underlying asset) until many investors had unexpected and seemingly unexplainable losses. The analysis offers an explanation for the confusion related to these products: the investors appear to have been ignorant of the outcome distribution's skew arising from high per-period volatility.

Chapter 2. The second chapter is concerned with the measurement of applicant quality on the German online peer-to-peer lending platform *smava.de*; specifically with the interaction of a loan applicant's gender and quality. In this context, it is investigated whether female loan applicants' success chances are more or less correlated with quality than males'.

In most discrimination studies (see, e.g., Bertrand and Mullainathan (2004), Pager (2003), Nunley et al. (2014), and Kaas and Manger (2012), among others) simple proxies for application quality, such as an additional reference letter or a positive criminal background check for job applications, are used with an often all too flexible interpretation. These proxies' statistical correlation with application success might indeed differ by gender, race, or any other group characteristic but it cannot be ruled out that this is rather due to selection effects or measurement error than evidence of discrimination. A substantial measurement problem arises in these studies simply because the objective of the potential employer is not self-evident.

This chapter's analysis of peer-to-peer online lending contributes to the discrimination literature by reasonably reducing the quality measure to a single number: a loan application's expected internal rate of return. The data set offers all characteristics of the loan application and of the applicant that are available to the potential lender, allowing for an assessment of this measure of quality in detail. The nature of the interaction between lender and borrower on the online lending platform precludes any other relation between them. Risk considerations are also minimal, due to the platform's specific insurance mechanism, implying that the expected rate of return is a natural candidate for the lender's objective.

Using this inferred measure of quality, funding success is analyzed with a particular focus on the interaction between gender and quality. Measurement error is addressed by modeling the applicant's quality in detail and by including statistical methods (the SIMEX procedure of Stefanski and Cook (1995)) to correct for measurement error in the measured quality. The results show that women have higher success rates than men, conditional on quality, but this gender difference is driven by a larger increase of men's success rate in quality: women appear to get the benefit of the doubt, such that low-quality applications of women are almost equally successful as high-quality applications of women and men. The low-quality applications of men, in contrast, are much less likely to be successful. These results are robust to a variety of specifications.

It is also shown that simpler proxies of quality are less precise proxies and yield different conclusions. One natural candidate proxy for quality is the applicant's offered loan rate. Its correlation with success, like that of the expected internal rate of return measure, suggests that women enjoy positive discrimination but are, if anything, harmed by offering higher quality. An alternative proxy, a loan applicant's credit rating, does only suggest a mild gender effect. In this sense, one can read this chapter as cautioning that the choice of proxies for quality is highly important for the conclusion.

Chapter 3. The third chapter presents a novel experimental design for eliciting subjective beliefs while accounting for the empirical fact that respondents usually cannot be relied upon to provide answers that satisfy the laws of probability. In large-scale surveys, for example, up to 60% of respondents violate the additivity axiom of Kolmogorov (1933) and its related concept of monotonicity (see, e.g., Delavande and Rohwedder (2008) and van Santen et al. (2012)). In order

to prevent the econometric problems of dealing with such data, the proposed approach confronts respondents with nothing more than binary choice questions. Binary comparisons require only that respondents are able to rank two objects, not that they are able to make probability statements or perform complicated operations with them. It is expected that answering binary choice questions may be much easier than stating subjective probabilities or even constructing confidence intervals, thus gaining an accurate elicitation of such beliefs in a user-friendly and timely manner.

The methodology is very general and can be applied to different domains (e.g., environmental variables like the weather or market prices like stock market returns) as well as probabilistic structures of almost arbitrary complexity. It can be used to estimate the relative likelihoods of discrete events or to estimate the entire probability distribution over a continuous state space as, indeed, is the case in the application described in this dissertation.

A stochastic choice model assumes that respondents have a subjective probability distribution from which they generate noisy binary judgments. Given enough binary comparisons and assumptions on the functional form of the probability distribution and the distribution of the noise term, the parameters of the underlying distribution can be estimated. The econometric method is agnostic about the size of the systematic component of responses and can accommodate and, in fact, identify anything from responses that are strictly consistent with the laws of probability to responses that are made entirely at random.

However, the psychological simplicity of the binary question design comes at a cost: binary choices, by their very nature, contain very little information. That is, binary events allow only for a (noisy) ordinal ordering. Therefore, the stochastic choice model is used not only to estimate the parameters of the distribution *ex post* but also to choose the questions asked *ex ante*. The questions are chosen such that their answers contain maximal information in a well-defined statistical sense and this is done adaptively: After any history of responses the model is estimated and the next question is chosen optimally *given the data*.

Model simulations show that despite the low informational content of binary choices this adaptive design allows for recovering the parameters of the choice model extremely well even with comparatively few judgments. Furthermore, results of an online study demonstrate the user-friendliness of the method and correlations of the elicited beliefs and other related measures that imply a satisfactory goodness of elicitation of the design.

Contents

List of Figures

List of Tables

1

Hidden Skewness

On the Difficulty of Multiplicative Compounding Under Random Shocks

This chapter is based on joint work with Olga Nottmeyer, Ludwig Ensthaler and Georg Weizsäcker.

1.1 Introduction

Many household investors face a particular mismatch in the time frames of asset return evaluations. They acquire their most important financial assets with the intention to liquidate them in the relatively distant future but the available return information concerns much shorter time intervals. Real estate investments, retirement savings plans or investments in college funds share this feature. In all of them, the relevant outcomes are the investments' performances over several decades but the available information concerns their short-term performances, like 1-year returns. To forecast the return on the planned (or any plausible) distant selling date, an investor needs to extract the price distribution at the selling date by compounding the available short-term return distributions. This is a formidable task for the average person.

Two biases may arise when forecasting the distribution of long-run growth. First, one may fail to compound the effects of multi-period growth, for a given growth rate. Second, one may ignore the skewness that arises over time and, e.g., confuse the mean return with the median return. The first of these biases, which we call "linearity bias" hereafter, has been studied predominantly in deterministic settings. For example, when asked to assess the total effect of accumulating 7% growth for ten periods, a substantial fraction of respondents gives an answer that is closer to 70% than to the actual 97%. The analyses of Stango and Zinman (2009) and Levy and Tasoff (2015) indicate that the bias is empirically relevant as it affects households' borrowing and saving decisions.[1] The effect is usually referred to as "exponential growth bias," but the main point of this paper is to extend the analysis of the exponential growth bias—appropriately defined—to stochastic settings and to demonstrate that it also includes the second bias. This second bias, which we refer to as "skewness neglect," is less well known in the academic literature[2] but investment practitioners and financial market regulators are aware of its effects (see Subsection 1.5). Investors apparently need to be made aware that the compounding of random growth can transform a symmetric 1-period return distribution into a skewed multi-period return distribution. An important real-world example of this is the family of leveraged exchange-traded funds (leveraged ETFs). These assets are highly volatile and have a fairly symmetric 1-period return distribution; holding them for multiple periods results in severe skew.

Our paper presents a series of incentivized laboratory experiments that extends the evidence on the perception of multiplicative growth to the stochastic domain and accounts for both of the above-described biases. As an example that demonstrates skewness neglect, consider the following stylized experiment. A very volatile asset either increases in value by 70% or decreases in value by 60% in every period, each growth rate realizing with a chance of one half. If the investor buys the asset she must hold it for twelve periods. With an initial value of 10,000, what would the asset likely be worth at the end of period 12? To ask this question in an incentive-compatible way, we let the participants bet on five possible outcome ranges for the period-12 value of the asset:[3] a) up to 6,400, b) between 6,400 and 12,800, c) between 12,800 and 19,200, d) between 19,200 and 25,600, or e) above 25,600. We then simulate the random process and if the simulated path ends up in the outcome range that a participant has bet on, she receives a prize of €20. If not, she receives nothing. The most popular answer is c), chosen by 43% of the participants, followed by d) (28%) and b) (17%). Response options a) and e) come tied bottom with a mere 6% of responses each. However, the optimal response is a); the median of the resulting distribution is 989 and the

[1] Both Stango and Zinman (2009) and Levy and Tasoff (2015) present survey evidence of a statistical connection between the bias and respondents' savings behaviors. Levy and Tasoff (2015) also analyse theoretical implications of the bias, e.g., an overestimation of future income that arises from too moderate time discounting of income. The effect can result in overconsumption if income is shifted to later time periods. Related effects are addressed in the experiments by McKenzie and Liersch (2011).

[2] The only other academic study that we are aware of is by Stutzer and Grant (2013), discussed in the next subsection.

[3] Appendix A.1.1 contains the details of the procedure. The experiment is not one of the main treatments of the paper.

probability that the process ends up in the lowest interval is 80%. A simple reasoning for this is that a value increase of 70% cannot recover a value decrease by 60%, hence most trajectories have a downward trend and the distribution is highly skewed already in period 12. The participants fail to realize this and instead report answers that are consistent with a confusion of mean and median. Their average expected payoff (based on their decisions) amounts to a meager €2 in this experiment, whereas the optimal response would earn them €16 in expectation.

Our series of experiments examines this kind of mistake systematically and finds that the participants' perception of stochastic growth deviates in predictable ways from the rational prediction. Both of the above biases are found to be relevant. Overall, the experimental results are in line with a simple model of misperceiving compound shocks. This model, which we label "exponential growth bias model," has the agent perceive growth as a linear process, in the sense that all multiplicative growth is mistaken as additive growth.[4] The model predicts both linearity bias and skewness neglect.

Importantly, the model also allows predictions about the strength of the two effects. It predicts that the agent has a fairly rational perception of the growth process in the case that both per-period volatility and per-period return are low. For larger volatility, skewness neglect becomes relevant and leads to an overestimation of the median. As the per-period return increases, the distribution of returns becomes more symmetric, the linearity bias becomes more dominant, and the agent underestimates the median. This somewhat intricate pattern of predictions cannot be generated by any of the biases alone but it is confirmed by the experimental data. Subsection 1.3.2 shows these effects in our main (novel) experiments, asking the participants to predict the most likely outcome of a growth process. In the binomial-tree processes that we use in this experiment, the most likely outcome is also the median and thus the responses can be used for assessing the subjectively perceived medians. All of our experiments are incentivized in ways that make truth-telling optimal irrespective of one's risk attitudes.

The experiments of Subsections 1.3.3, 1.4 and 1.5, and those in the Appendix A.1, go on to test the predictions of the model in different variations of the experimental setting. We vary the incentive schemes, the level of feedback as well as the nature of both the investment strategy and the underlying asset that the participants are asked to assess. Thereby, we can investigate the robustness of the effects and we can also inquire about several other implications of the exponential growth bias model. Qualitatively, almost all predictions of the model are borne out in the data, and often with large discrepancies to the rational prediction. For example, in treatments with high return volatility, about 90% of the participants overestimate the median. The model is also fairly successful in predicting the participants' misperceptions of the 10th and 90th percentile of the relevant long-run distributions: the 90-10 spread is generally underestimated. This holds both for binomial-tree assets and also for a class of more realistic assets that we base on the historical returns of the German DAX index. In the latter set of experiments, we model assets that emulate leveraged ETFs where the DAX index is the underlying asset. Such leveraged products have been popular with household investors in the U.S. in recent years (though with a different underlying asset) until many investors made unexpected and seemingly unexplainable losses. Our analysis offers an explanation for the confusion related to these products: the investors appear to have been ignorant of the skew arising from high per-period volatility.

The rest of this paper is organized as follows. Subsection 1.2 briefly discusses related literatures. Subsection 1.3 (Study 1) introduces the exponential growth bias model and the main experimental design. In Subsection 1.3.2 we report four experimental treatments that investigate the perceived medians and produce the above-described data pattern of linearity bias and skewness neglect, while Subsection 1.3.3 and the Appendix A.1.2 cover several variations of the elicitation method. Subsection 1.4 describes Study 2, which varies the format of Subsection 1.3.3 to investigate the perceived 10th and 90th percentiles. This study also varies the investment horizon. Subsection 1.5 (Study 3) reports on the extension to leveraged ETFs, and Subsection 1.6 concludes.

[4] The model is akin to that of Levy and Tasoff (2015) albeit developed independently.

1.2 Review of Related Literature

Classic studies in cognitive psychology discuss quite extensively to what degree the human cognitive apparatus is able to account for the distinction of linear versus nonlinear relations between variables. Wagenaar and Sagaria (1975) ask participants to predict an exponential data series representing an index for pollution. They find that participants strongly underestimate exponential growth. Wagenaar and Timmers (1978) show that this linearity bias is robust to the amount of information available to the participants and Wagenaar and Timmers (1979) demonstrate robustness of the effect to the framing of the information. Kemp (1984) surveys perceptions of changes in the cost of living. Respondents systematically underestimate the increase in cost, which is also in line with linearity bias. Much of the early data analysis uses responses to quiz-type questions, but a subsequent specialization of this literature more and more focuses on economic contexts, like the perception of compound growth from interest or loan payments. Eisenstein and Hoch (2005), Stango and Zinman (2009), Christandl and Fetchenhauer (2009), McKenzie and Liersch (2011) and Levy and Tasoff (2015) document that participants underappreciate the effects of compound interest and thereby predictably underestimate the compound effect of growth. Chen and Rao (2007) show that retailers can strategically use this bias by posting double dip price discounts (a discount of 20% followed by another 25% discount is perceived to be a 45% reduction, not the actual 40%). As described in the Introduction, our paper can be viewed as an extension of this literature to non-deterministic growth processes.

An important predecessor of our paper is the study by Benartzi and Thaler (1999) who, among other things, study biases in the compounding of long term distributions from a given short term distribution. Their experimental participants choose different hypothetical retirement plans depending on whether they observe the historical return distribution of retirement plans for a 1-year period or a 30-year period. Benartzi and Thaler (1999) relate this bias to the effects of myopic loss aversion (see also Samuelson (1963), Redelmeier and Tversky (1992), Gneezy and Potters (1997), and Klos and Weber (2005)). While we agree that myopic loss aversion likely plays a role in households' long term investment decisions, our experiments suggest that household decisions can also be misguided by a biased perception of the underlying growth processes.[5] This is also consistent with the only experimental paper on skewness neglect that we found, by Stutzer and Grant (2013). Their hypothetical investment experiments find an inflated investment rate in treatments where their participants have to calculate the compound return by themselves.[6]

Another related literature studies whether experimental participants have a correct understanding of financial options. We refer the reader to Gneezy (1996) and Abbink and Rockenbach (2006) for previous results in this—surprisingly small—literature. We note that the assets that we use in Subsections 1.3 and 1.4 have the same structure as the underlying asset in the well-known Cox et al. (1979) model of European call options. A consistent finding of misperceptions of such assets may therefore indicate a potential mispricing. This is not further studied in our paper, which focuses on the underlying asset itself.

[5] A distinction between our study and the existing experimental work on myopic loss aversion is that the existing papers largely make use of additive growth processes.

[6] The experiment by Stutzer and Grant (2013) uses a quite similar experimental wording as the experiment described in Subsection 1.3.3 and in the first version of this paper (Ensthaler et al., 2010), despite having been developed and written independently. A separate and important experimental literature examines the preferences regarding skewness, see Deck and Schlesinger (2010), Brünner et al. (2011), Ebert and Wiesen (2011), and Eckel and Grossman (2014). We restrict this paper to the perception of the distribution, not its valuation.

1.3 Study 1: Assessments of Median and Mode

Study 1 is composed of the two partial studies 1(a) and 1(b), each testing participants' perceptions of an asset's mode and/or median. We consider one of the most elementary assets covered in the finance literature: the binomial-tree asset with fixed maturity (Cox et al., 1979). The multiplicative per-period growth μ_t of this asset is a binary random variable $\mu_t \in \{\mu^h, \mu^l\}$, where the percental uptick $\mu^h \geq 0$ and the percental downtick $\mu^l \geq 0$ are equiprobable in each period $t = 1, ..., T$.

1.3.1 The Exponential Growth Bias

We start the analysis by presenting a simple model of biased decision making. Consider a decision maker who ignores compounding of interest: when asked to predict the accumulated value gain of an investment that yields a per-period interest of r over T periods, she quotes a total gain of rT. That is, she wrongly perceives the absolute changes, not the relative changes, to be constant across the periods. This feature is the sole bias of our model—the exponential growth bias—and we can readily extend it to the domain of stochastic growth.

Formally, let Y_0 denote the known initial price of an asset with a random price series $\{Y_0, Y_1, ...\}$ and let μ_t be the random variable describing the relative price changes occurring in t, e.g., $Y_1 = Y_0\mu_1$. An unbiased decision maker correctly perceives the true distribution of the period-T price as $Y_T = Y_0 \prod_{t=1}^{T} \mu_t$. In contrast, an exponential growth biased (EGB) decision maker perceives the price in t as $\tilde{Y}_T = Y_0(1 + \sum_{t=1}^{T} (\mu_t - 1))$. That is, for each t she perceives the absolute difference $Y_t - Y_{t-1}$ to be given by t's growth rate applied to the initial value Y_0. As a result, the EGB decision maker misses out on all effects of multiplicative compounding, which may or may not occur in the true growth process.

To investigate the effects of the bias, Study 1 and Study 2 consider binomial-tree price series with a constant distribution of relative price growth μ_t (but with distributions of absolute differences that vary across time, which is ignored by the EGB decision maker). Here, while the actual distribution of Y_t is skewed and approaches a lognormal distribution for large t, the EGB decision maker perceives a binomial distribution that is symmetric with its mean being equal to the median and mode: symmetry is preserved under addition of random variables.[7] The EGB model thus predicts full skewness neglect. In particular, one can check that with a strictly positive per-period average growth $\mathbb{E}[\mu] > 0$ and under the condition that $\mu^h \mu^l < 1$, the EGB decision maker overestimates the median for $t > 1$. Under the same conditions, the model also predicts a directed linearity bias: the EGB decision maker underestimates the mean, for $t > 1$. One can also check that if both μ^h and μ^l are increased by the same amount Δ, then the distribution of Y_t becomes more and more symmetric, so that the EGB decision maker's skewness neglect becomes less and less important.

In the following, we present our experiments that test these qualitative (directed) predictions of biased decision making, in each case generated by simple numerical applications of the EGB model.[8] Our main empirical focus lies on measuring the perception of the median of Y_T. We also exploit the fact that for binomial-tree processes the median is identical to the mode of Y_T, to which both optimal and EGB decision makers agree (despite disagreeing on the value). This allows formulating alternative elicitation tasks, equivalently asking for median or mode.

[7] This error could also be interpreted as the decision maker wrongly computing the arithmetic mean over returns when calculating the median instead of working with log-returns.

[8] We also discuss the model's point predictions for completeness; but as a model of such simplicity cannot plausibly capture the precise decision process we focus our statistical analysis on the qualitative predictions.

1.3.2 Study 1(a): Biased Perception of the Mode

1.3.2.1 Experimental Design

Participants in Study 1(a) are presented with a security whose price is currently at $Y_0 = 100$ and changes by a factor $\mu_t \in \{\mu^h, \mu^l\}$ with equal probabilities during each period and with all random draws being independent. The participants' task is to locate the mode of the security's outcome distribution after $T = 12$ periods. The task is made incentive compatible as follows. After a participant's response, the experimenter simulates a set of 100 values of Y_T. If at least one of these simulated values differs by less than 1 from the participant's stated value she receives a bonus of €20, otherwise not. The procedure thus prompts the participant to report the location (more precisely, an interval of length 2) where she perceives Y_T's highest likelihood. The optimal response would be to report the mode of Y_T. Notice that reporting the mode is optimal irrespective of risk preferences and of the nature of the perceived Y_T: the incentive scheme uses only two possible payments—receive a bonus versus not—making it optimal for any participant with monotonic preferences to maximize the subjectively perceived probability of receiving the bonus by stating the price that she thinks is most likely. The procedure is also simple to understand and allows asking a straightforward question about the participants' prediction of the price evolution of the asset.[9]

Each participant is asked to report a prediction on two different securities in order to increase the number of observations and thereby the power of our statistical tests. One of the two responses is randomly picked to be payoff relevant at the conclusion of the experiment.

Overall, Study 1(a) covers four securities that only differ in the values $\{\mu^h, \mu^l\}$, appearing pairwise in two treatments. Participants in treatment 1 assess the modal values of Security 1 ($\mu^h = 1.7, \mu^l = 0.4$) and Security 2 ($\mu^h = 1.075, \mu^l = 1.025$), whereas participants in treatment 2 assess the modal values of Security 3 ($\mu^h = 1.4, \mu^l = 0.7$) and Security 4 ($\mu^h = 1.8, \mu^l = 1.1$). Each participant thus faces one security which can depreciate as well as appreciate, and one security which can only appreciate. In treatment 1, the two securities have identical means but different per-period volatilities (as measured by the spread $(\mu^h - \mu^l)$) and in treatment 2, the two securities have different means but identical per-period volatilities. Moreover, the mean of Security 3 is identical to that of Security 1 and 2.

Participants are randomly assigned to treatments 1 or 2. To account for possible learning effects, the order of the two securities randomly varies between the participants within a treatment. All 127 participants (63 in treatment 1 and 64 in treatment 2) are students at Technical University Berlin. Six sessions, three in each treatment, are conducted in a computer-based format using the software z-Tree (Fischbacher, 2007). Participants receive a participation fee of €5 in addition to their possible bonus of €20.

1.3.2.2 Exponential Growth Bias Prediction

The securities in Study 1(a) are specified such that they allow predictions about the relative strengths of the above-described two effects, linearity bias and skewness neglect. To examine the effect of a large per-period volatility, we consider Security 1 ($\mu^h = 1.7, \mu^l = 0.4$) with +70% and -60% as possible percentage changes and predict a strong effect of skewness neglect. With

[9] The procedure is novel to the experimental literature, to our knowledge. All experimental instructions can be found in the online appendix (available as supplemental material at `https://drive.google.com/open?id=0B4TlJkDn0R5LazY1a0RRS2ljeE0`), including the instructions for an understanding test that participants had to pass.

a perceived constant distribution of absolute changes that lie in $\{-60; +70\}$, the EGB decision maker perceives a symmetric distribution of the period-12 price that locates mode, mean, and median at $\mathbb{E}(\tilde{Y}_T) = 160$. Thus, skewness neglect leads to a strong overestimation of the true mode of Security 1, which is at 9.89. This effect stems entirely from skewness neglect, whereas linearity bias has only a mild effect: the EGB decision maker's belief about the mean is close to the true mean $\mathbb{E}(Y_T) = 179.59$.

For a relatively lower per-period volatility, as captured by Security 3 ($\mu^h = 1.4, \mu^l = 0.7$) with +40% and -30% as possible percentage changes, skewness neglect becomes less extreme and results in a weaker, but still sizable, overestimation of the mode: the EGB model predicts a response of 160 instead of the correct 88.58.

Further decreasing per-period volatility, as for Security 2 ($\mu^h = 1.075, \mu^l = 1.025$), results in the predictions that the EGB decision maker has a fairly rational perception of the growth process: she perceives the most likely period-12 price of Security 2 at 160 instead of the correct 178.97. That is, even a model allowing for both skewness neglect and linearity bias is fairly ineffective here and predicts a mild underestimation of Security 2's mode.

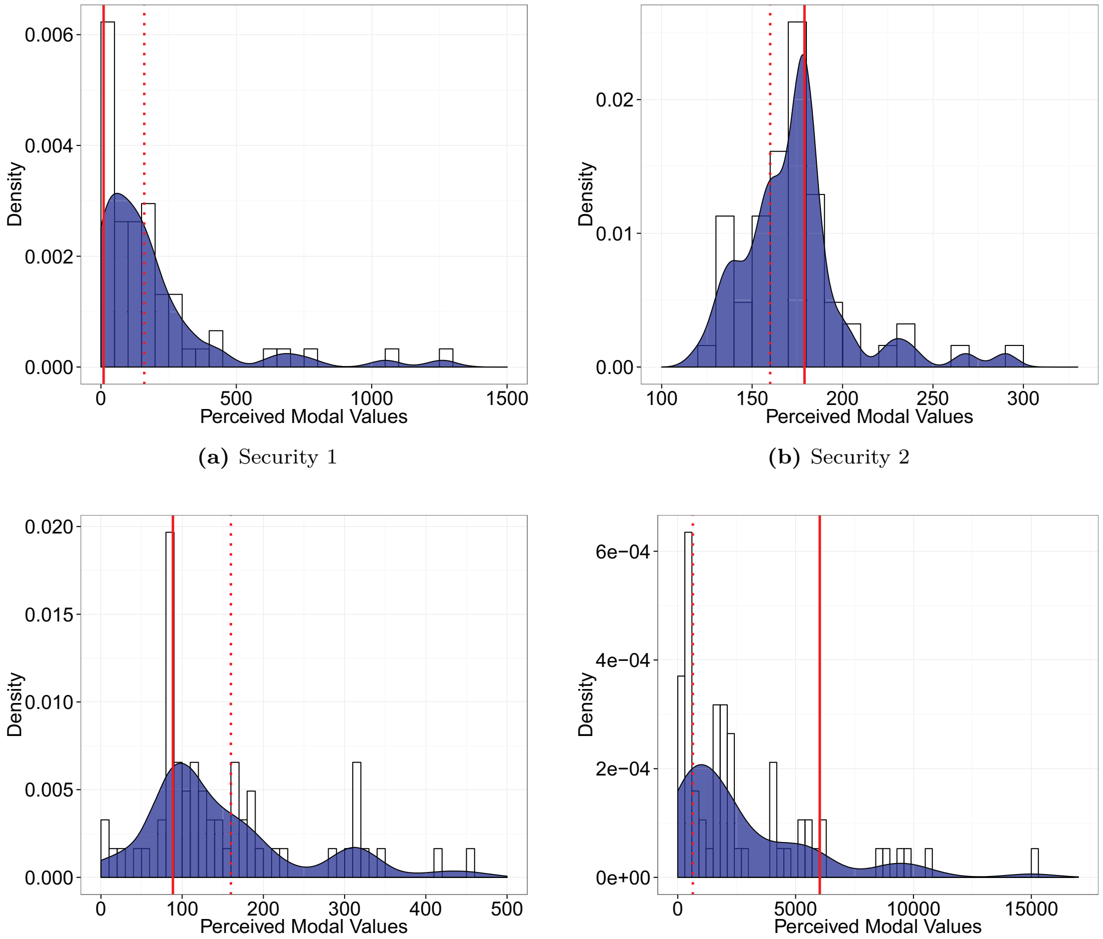
Security 4 ($\mu^h = 1.8, \mu^l = 1.1$) allows for a much stronger effect towards underestimation, which is due to linearity bias. Here, the EGB decision maker perceives the most likely period-12 price at 640, while the true mode of the price distribution lies at 6,025.47. With such high per-period mean growth, skewness considerations become less important than the effect of linearity bias.

1.3.2.3 Results

Figure 1.1 illustrates the distributions of subjective mode perceptions for all four securities estimated from our experimental data. The participants' predictions for Security 1 ($\mu^h = 1.7, \mu^l = 0.4$) are displayed in Figure 1.1(a) and show a substantial degree of overestimation (i.e., most of the probability mass is located to the right of the vertical solid line, indicating the optimal response). Consistent with skewness neglect, 87% of the participants overestimate the true mode and this frequency of overestimation lies significantly above 50% (p-value<0.001, one-sided binomial test). Although the data show a peak in the neighborhood around the optimal value, most participants' degree of overestimation is substantial. Half of them predict the mode of the distribution to lie above 120—more than 12 times the true value.

Figure 1.1(c) illustrates the subjective mode perceptions for Security 3 ($\mu^h = 1.4, \mu^l = 0.7$). As for Security 1, the data show a notable proportion of participants, 70%, overestimating the mode. While the frequency of overestimation lies significantly below that of Security 1 (p-value<0.001, two-sample binomial test), it is still significantly greater than 50% (p-value<0.001, one-sided binomial test). These observations are consistent with the EGB model in the sense that the model predicts an overestimation for both securities and a larger overestimation for Security 1 than for Security 3. But also apart from the model, the comparison between Security 1 and Security 3 is relevant as it shows the effect of skewness neglect in isolation: the mean is constant between them whereas the higher volatility in Security 1 changes median and mode. The higher frequency of overestimation in Security 1 illustrates that participants do not fully appreciate this difference.

The participants' assessments of Security 2 ($\mu^h = 1.075, \mu^l = 1.025$) are depicted in Figure 1.1(b) and (again consistent with the EGB model) show a different picture. Only a minor underestimation for Security 2 appears: 57% of participants state modal values below the true mode, a proportion that is not significantly greater than 50% (p-value>0.15, one-sided binomial test). Moreover, the median response is not significantly different from the optimal value (p-value>0.05, Wilcoxon signed-rank test). Neither skewness neglect nor linearity bias show to be relevant for this security.



(a) Security 1

(b) Security 2

(c) Security 3

(d) Security 4

Figure 1.1: Densities of the subjectively perceived modal values for securities 1 through 4. Solid lines indicate rational benchmarks at 9.89 (Security 1), 178.97 (Security 2), 88.58 (Security 3), and 6,025.47 (Security 4). Dotted lines illustrate EGB predictions for mean, mode and median.

The perceptions for Security 4 ($\mu^h = 1.8, \mu^l = 1.1$), with a higher average per-period return, are illustrated in Figure 1.1(d) and show a substantial degree of underestimation. Here, 89% of the participants state responses that lie below the true mode. This share is significantly larger than 50% (p-value<0.001, one-sided binomial test) and also significantly larger than the share of participants who underestimate Security 2's modal value (p-value<0.001, two-sample binomial test). Once again, these observations are consistent with the much stronger prediction of the EGB model for Security 4 than for Security 2. Moreover, it is notable that the data confirm the EGB model's prediction that linearity bias is more relevant than skewness neglect in Security 4.

1.3.3 Study 1(b), Robustness Checks

Study 1(b) focuses on eliciting the median. We use a novel choice list mechanism to identify bounds on the median of each participant's subjectively expected price distribution of a binomial-tree asset. As for Study 1(a), Study 1(b) ensures incentive compatibility under a wide set of preferences by using only two possible payments per choice problem—receive a bonus versus not. We also let the participants repeat this task over five rounds.[10]

1.3.3.1 Experimental Design

	Thresholds for Security A	Thresholds for Security B	Your decision (A or B)
Task 1	100	10,000	_
Task 2	500	10,000	_
Task 3	2,000	10,000	_
Task 4	6,000	10,000	_
Task 5	9,000	10,000	_
Task 6	12,000	10,000	_
Task 7	20,000	10,000	_
Task 8	35,000	10,000	_
Task 9	90,000	10,000	_
Task 10	250,000	10,000	_

Table 1.1: The 10 binary choices.

In each round of the experiment, two risky securities are on offer and the selling price of the chosen security determines whether or not the participant receives the bonus. Security A follows a binomial-tree -60%/+70% process over 12 periods that is identical to Security 1 in Study 1(a) with the sole exception that its initial price now is 10,000. A participant who chooses this security receives the bonus if the selling price at maturity exceeds a given threshold t_A. The alternative choice is Security B, which yields the bonus with probability one half. One can immediately see that it is subjectively optimal for a participant to choose Security A if and only if she believes that Security A yields the bonus with a probability more than one half. A choice for Security A thus reveals that the median of her subjective probability distribution of Security A's selling price is above t_A.

For a balanced experimental design we describe Security B analogously to Security A, with the difference that Security B has only a single equiprobable price change of -60% or +70% during the 12 periods. A participant who chooses Security B receives the bonus if the selling price of Security B exceeds a separate threshold t_B. This threshold is fixed at the initial price of 10,000 throughout the experiment (hence Security B holds a 50-50 chance of receiving the bonus) whereas the threshold t_A varies between 10 different values. Each participant makes a choice between Security A and B for each of the 10 values of t_A, allowing us to infer bounds on her subjective median of Security A's selling price distribution.

Table 1.1 lists the 10 choice problems (Task 1, Task 2, etc.) as seen by the participants. Given that the true median of Security A's selling price distribution is 989, the rational prediction is for the participants to choose Security A in Task 1 and Task 2 and to choose Security B in all subsequent tasks. After the participants make their 10 choices, they receive individual feedback in the form of a sample pair of selling prices of Security A and B. This concludes the first round of the experiment. The experiment is then repeated for four additional rounds of the same nature, each including 10 choices and individual feedback.[11] Three sessions are conducted in a paper-and-pencil format, with

[10] In a further treatment variation of Study 1(b), we additionally provide the participants with an explicit calculation of the distribution of compound price changes after two periods for the respective security and we point out the asymmetry in the price distribution. The observed choice bias decreases strongly in this treatment, consistent with the presumption that the bias stems from a cognitive problem and is not driven by the particular choice format. A detailed description of this treatment is in Appendix A.1.2.

[11] Each additional round comes with the chance to earn a new bonus but this does not affect the simple optimality conditions for choice. Independent of other choices it remains optimal to choose A iff the subjective median is above t_A.

68 student participants at Technical University Berlin. Participants receive a participation fee of €5 and a possible bonus of €5 per round. That is, participants can earn up to five bonuses of €5 each, one per round of the experiment. After completing all choices, each participant receives five random draws of integers between 1 and 10 to determine which of the 10 choice problems in each round is payoff relevant for her.

1.3.3.2 Exponential Growth Bias Prediction

The EGB model predicts that a biased decision maker perceives a binomial distribution with mean equal to mode equal to median at 16,000. She would therefore overestimate the true median (989) by an order of magnitude and choose a switching value in the interval [12,000; 20,000).

For notation, let $q_{0.5,i}$ be the elicited *lower* bound of participant i's assessment of the median: i invests in Security A for all values $t_A \leq q_{0.5,i}$ and invests in Security B for all strictly larger t_A. For the sake of simplicity we restrict attention to cases where participants' choices reveal such a unique switching value, a property that is true in 93% of our data.[12] By analogy, let $q_{0.5}$ be the rational benchmark for $q_{0.5,i}$ (dropping the subscript i), i.e., the lower bound of the median that would be elicited from a decision maker who behaves optimally. Here and elsewhere in the paper, we focus on revealed lower bounds when applicable.

1.3.3.3 Results

Table 1.2 lists the frequencies with which the participants' subjective medians lie in relevant ranges of Security A's selling price distribution, over the five rounds. In round 1, not a single participant gives the optimal response of $q_{0.5} = 500$ (i.e., optimal switching at Task 3). Instead, 98% of the participants reveal that their subjective medians are above 2,000. The results of rounds 2 to 5 show that a proportion of 86% of participants overestimate the median still in round 5. (The proportion lies significantly above 50%, with p-value<0.001 in a one-sided binomial test). The modal choice in round 1 (41% of participants) indicates a subjective median between 9,000 and 12,000, with the next-higher interval [12,000; 20,000) attracting 20% of participants' choices.

Range of subjective median	Share of participants switching from A to B				
	Round 1	Round 2	Round 3	Round 4	Round 5
[0 ; 100)	0.018	0.000	0.000	0.018	0.000
[100 ; 500)	0.000	0.000	0.000	0.000	0.035
[500 ; 2,000)	0.000	0.054	0.072	0.072	0.107
[2,000 ; 6,000)	0.036	0.145	0.127	0.200	0.303
[6,000 ; 9,000)	0.107	0.090	0.254	0.309	0.142
[9,000 ; 12,000)	0.411	0.381	0.309	0.236	0.196
[12,000 ; 20,000)	0.196	0.181	0.109	0.127	0.142
[20,000 ; 35,000)	0.179	0.090	0.109	0.036	0.053
[35,000 ; 90,000)	0.054	0.054	0.000	0.000	0.017
[90,000 ; 250,000)	0.000	0.000	0.000	0.000	0.000
[250,000 ; ∞)	0.000	0.000	0.018	0.000	0.000

Table 1.2: Subjective median ranges over the five rounds.

[12] If a participant has multiple switching points in one round, her answers in the remaining rounds are still considered in our data analysis. None of our conclusions would change if we dropped all responses by participants switching more than once in at least one round (12% of participants), or if we included all data and considered each of the 10 tasks separately.

1.4 Study 2: Additional Quantiles

In this subsection we examine the robustness of the exponential growth bias predictions with respect to variations of the investment horizon and the choice of quantiles of the subjective distributions that we elicit. The results confirm the EGB model's implication that the compound distribution is perceived as too symmetric and too narrow-band if there is substantial randomness in the growth process.

1.4.1 Experimental Design

There are two treatments in Study 2, both involving assets similar to those in Study 1. Participants can buy a Security A at a price of 100. If they buy it they have to sell it after T^k periods, where k indexes the treatment. The price of Security A moves by about +/-20% in each period: in both treatments, *High Volatility Short* (HVS) and *High Volatility Long* (HVL), the parameters specifying upticks and downticks are $\mu^{h,HVS} = \mu^{h,HVL} = 1.212$ and $\mu^{l,HVS} = \mu^{l,HVL} = 0.811$. The sole difference between these two treatments is in the length of time until maturity: $T^{HVS} = 14$ and $T^{HVL} = 140$.[13]

As in Study 1(b), a participant of Study 2 who buys Security A receives a fixed bonus if the selling price at maturity exceeds a given threshold t_A. These thresholds differ between treatments and are listed in Table 1.3. The alternative choice option is Security B which yields the bonus with a certain probability.[14] To elicit three different subjective quantiles of Security A's selling price distribution, Security B has three different specifications. Security B1 yields the bonus with 90% probability, B2 with 50% and B3 with 10%. Accordingly, each participant faces three choice lists. First, she chooses between Security A and B1 for the different thresholds of Security A. This allows us to infer bounds on her subjective 10th percentile of Security A's selling price distribution. For example, suppose that participant i in treatment HVS chooses Security A over Security B1 in Task 1 and Task 2 and chooses Security B1 over Security A in tasks 3 through 10. Inspecting Table 1.3 (first column) we see that this is subjectively optimal iff participant i's subjective 10th percentile for Security A's selling price distribution is between 30 and 45. In line with our previous notation, we would thus record the elicited lower bound of i's subjective 10th percentile for Security A's selling price distribution as $q^{HVS}_{0.1,i} = 30$.

As her second set of tasks a participant faces the analogous choices between Security A and B2 (with the same list of thresholds for Security A). This allows us to infer a lower bound on her subjective median of the selling price distribution, $q^{HVS}_{0.5,i}$. Finally, she faces the analogous list of choices between Security A and B3, allowing us to infer a lower bound on her subjective 90th percentile of the same price distribution, $q^{HVS}_{0.9,i}$.[15]

After the participants make their 30 choices, the computer terminals report feedback to them in the form of a sample selling price of Security A. In each treatment, this concludes the first round of

[13] In three further treatments, the price motion is approximately deterministic (i.e., the price volatility is very low) and the price has positive growth with certainty. A detailed description and the results can be found in Appendix A.1.3.

[14] Different from Study 1(b), the instructions simply report to the participants the probability with which Security B yields the bonus, without referring to a separate threshold t_B.

[15] After the elicitation of the subjective quantiles we also ask for the participants' beliefs of Security A making a profit. We do not use the resulting data in the analysis but refer to the paper's previous version (Ensthaler et al., 2013) and to the instructions for a description of the experimental details and the results. Appendix A.1.5 provides an instruction sample of a related treatment.

the experiment. The experiment is then repeated for four additional rounds.[16] All 58 participants are undergraduate students at University College London. Each of the two treatment conditions is faced by a random subset of participants in each session, without making them aware that other participants face different treatment conditions (N=29 in both HVS and HVL). Participants receive a participation fee of £5 and a possible bonus of £5 per round. In each round, a single choice is randomly determined to be payoff-relevant, giving an ex-ante incentive to act optimally in each task.[17] For a simpler data analysis, the participants' computer interfaces restrict responses to satisfy two constraints. First, responses must exhibit at most one switching point on a choice list between Security A and a single B-type security. That is, a participant cannot switch back and forth between Security A and the respective B-type security. Second, the elicited quantiles must be ordered in a consistent way: a participant cannot switch from Security A to B1 at a threshold that exceeds the threshold at which she switches from Security A to B2, which in turn cannot exceed the threshold at which she switches from Security A to B3.[18]

	Values of t_A	Values of t_A
	HVS	*HVL*
Task 1	15	2
Task 2	30	5
Task 3	45	15
Task 4	65	60
Task 5	95	140
Task 6	125	230
Task 7	155	350
Task 8	190	550
Task 9	225	700
Task 10	265	1,000

Table 1.3: The thresholds t_A by treatment condition.

1.4.2 Exponential Growth Bias Prediction

The treatment comparison in Study 2 focuses on the effects of investment horizon variations. With a subjective constant distribution of absolute changes in $\{-18.9; 21.2\}$, the EGB decision maker perceives a perfectly symmetric selling price distribution with its median at 116.10 in treatment HVS and at 261.00 in treatment HVL. The EGB model thus predicts that participants overestimate the true median (88.64 in HVS and 29.96 in HVL). This effect is rather mild in HVS and much more pronounced in HVL.

Moreover, as discussed in Subsection 1.3.1, the EGB decision maker perceives no skewness in the distribution of the selling price of Security A. The true distribution Y_T is skewed, however, and the EGB model thus predicts a false assessment of the 10th and 90th percentiles. In particular, the EGB decision maker fails to realize that the right tail of the distribution is long, especially with a long investment horizon. We chose the experimental parameters such that under the EGB model this bias would have no discernible effect in treatment HVS, but predicts an effect in HVL. That is, the model predicts that the 90-10 spread is too narrow-band in treatment HVL.[19]

[16] All participants passed an understanding test, in a few cases after asking for some additional explanations. In contrast to Study 1, participants in Study 2 are supplied with a hand-held calculator that they can use throughout the experiment.

[17] Participants can earn the bonus either through the quantile elicitation task or through the profit probability elicitation task (see Footnote 15). For each round and each participant, the relevant task type (quantile or profit probability) is determined by a simulated coin flip at the end of the experiment.

[18] The instructions explain that violations of these constraints are subjectively suboptimal and the experimental software shows an error message if a participant violates either of the two constraints. Only 2% of the participants' inputs receive one or more error messages.

[19] The point predictions of the EGB model for the 10th and 90th percentile for treatment HVS are 35.90 and 196.30, respectively. The rational quantile assessments are only marginally different at 39.69 and 197.98. In treatment

1.4.3 Results

We start the analysis of the results with a descriptive overview. We then present interval regressions that impose a normal decision error and allow estimating the underlying quantiles while taking into account the discrete nature of the experimental data.

1.4.3.1 Descriptive Overview

In treatment HVS (with the 14-period time horizon), at least half of the participants within each round strictly overestimate the median of the stochastic process, i.e., switch at least one task later than rational. Precisely, the shares of participants that reveal such a misperception of the median are: {Round 1: 55%, Round 2: 65%, Round 3: 55%, Round 4: 69%, Round 5: 62%}. However, only the share in Round 4 is significantly greater than 50% (p-value<0.05, one-sided binomial test), which is consistent with the EGB model's prediction of a mild median misperception for HVS.

As also predicted by the EGB model, the 90-10 spread perceptions in HVS, too, deviate only slightly from the rational prediction. This pattern appears in each of the five rounds. The respective shares of participants that underestimate the spread are as follows: {Round 1: 65%, Round 2: 65%, Round 3: 72%, Round 4: 72%, Round 5: 65%}. These shares are significantly greater than 50% (p-values<0.05, one-sided binomial test) in Round 3 and Round 4 only.

We get the same qualitative results for treatment HVL (with the 140-period investment horizon), but they are much stronger. This is further support for the EGB model. More than 75% of participants overestimate the median in each round of this treatment: {Round 1: 93%, Round 2: 79%, Round 3: 79%, Round 4: 79%, Round 5: 79%} with all shares exceeding 50% significantly (p-values<0.001, one-sided binomial test). Moreover, for all but two rounds, these shares are significantly greater in HVL than in HVS (p-values<0.05, two-sample binomial test).

Finally, as the EGB model predicts, we observe that the perceived 90-10 spread in HVL is underestimated by almost all participants in all rounds: {Round 1: 89%, Round 2: 96%, Round 3: 100%, Round 4: 96%, Round 5: 93%} (p-values<0.001, one-sided binomial test for >50%).[20] For each round these shares are significantly greater in HVL than in HVS (p-values<0.05, two-sample binomial test).

1.4.3.2 Interval Regressions

The above descriptive analysis is based on the elicited lower bounds of the participants' perceived quantiles of Security A's price distribution. This subsection investigates point estimates instead of lower bounds and employs interval regressions—a modified version of the ordered probit regression (see e.g., Wooldridge (2002)). The analysis takes into account the interval nature of the data and assumes that the subjectively perceived quantiles are subject to normally distributed disturbances. Under this assumption, the mean of the participants' subjective quantiles can be estimated

HVL, the EGB model predicts the 10th percentile at 0 and the 90th percentile at 581.80, with the corresponding rational values at 1.20 and 745.58.

[20] EGB predictions and rational benchmarks for the 10th and 90th percentiles differ only for the latter in HVL. There, the shares of those underestimating the 90th percentile range between 72% and 96% over the rounds and are always significant, supporting the EGB model (p-values<0.05, one-sided binomial test for >50%). In HVS, where EGB and rational predictions coincide for the 90th percentiles, the respective shares are much lower and not significantly different from 50% (ranging from 31% to 58%).

via maximum likelihood, and standard hypothesis testing applies. Figures 1.2 and 1.3 report the corresponding estimates of the population means for the subjectively perceived quantiles of Security A's selling price distribution, separately for each of the two treatments and for each round. The horizontal dashed lines depict the benchmark rational predictions for the respective treatment-specific quantiles. The point estimates of the participants' average perceptions are enclosed by 95% confidence intervals.[21]

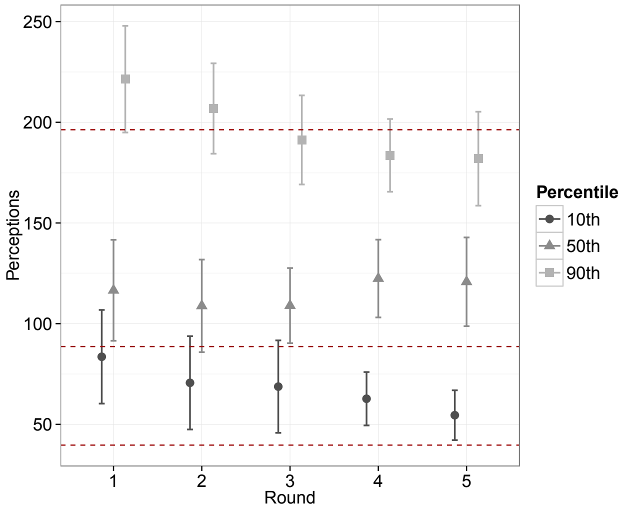


Figure 1.2: Point estimates of the participants' subjective quantiles of Security A's selling price distribution in HVS enclosed by 95% confidence intervals. Dashed lines indicate rational benchmarks for the 10th percentile (lowest), median (middle) and 90th percentile (uppermost).

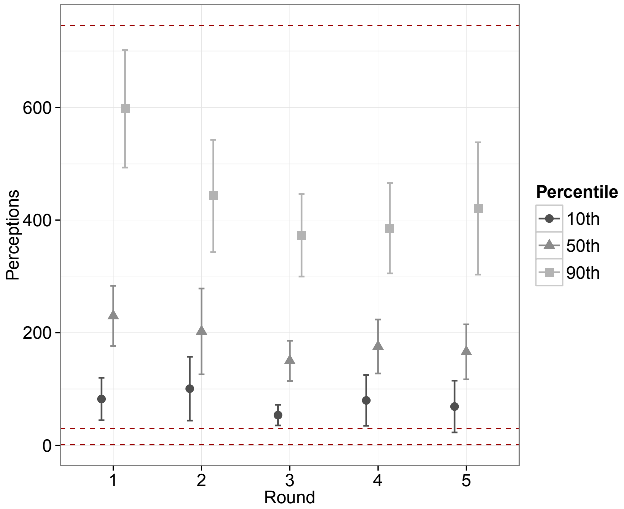


Figure 1.3: Point estimates of the participants' subjective quantiles of Security A's selling price distribution in HVL enclosed by 95% confidence intervals. Dashed lines indicate rational benchmarks for the 10th percentile (lowest), median (middle) and 90th percentile (uppermost).

Figure 1.2 confirms the previous subsection's descriptive analysis. In HVS, the round-wise 95% confidence intervals of the point estimates for the average median perceptions (triangle) are only

[21] For a detailed listing of interval regression estimates, see Appendix A.1.4.

marginally above the rational predictions (middle dashed line). In contrast to that and in accordance with the EGB model, the HVL median perception confidence intervals (Figure 1.3) differ significantly from the rational benchmarks for all five rounds.

Regarding the perceived skewness, the estimates also confirm the EGB model, as they indicate that on average the underlying subjective distributions are only marginally more symmetric than the rational prediction in HVS (Figure 1.2) but significantly more symmetric compared to the rational predictions in HVL (Figure 1.3). Once again, we observe that this pattern is robust over the rounds. The same is true for average perceptions of spread. The difference between the estimates for 10th percentile perceptions (circle) and 90th percentile perceptions (square) indicate subjective distributions that are much more narrow-band compared to the rational predictions (lowest and uppermost dashed lines) in HVL than in HVS.

1.5 Study 3: EGB in the Perception of ETFs

In this study we test the robustness of the EGB model in a setting where the asset price depends on real-world data. We simulate leveraged and unleveraged exchange-traded funds (ETFs) on past data of the German stock market index DAX30 to examine how changes in volatility affect participants' perceptions of real-life growth processes.

Leveraged ETFs move by a given multiple relative to an underlying asset, compounded at the end of each trading day. A triple leveraged ETF on the DAX30 index, for example, changes by +3% on a trading day if the DAX30 increases by 1% on that day and it changes by -3% if the DAX30 falls by 1%. Leveraged ETFs are a popular asset class amongst household investors but have come under severe scrutiny as many investors were perplexed when the products made a loss in a period where the underlying index made a gain.[22] Our experiment confirms the prediction derived from the EGB model that skewness and spread are strongly underestimated if the volatility is high.

1.5.1 Experimental Design

There are two treatments in this study. In both treatments, Security A is an ETF based on the DAX30. The two treatments differ only in that their respective versions of Security A differ in per-period volatility. In treatment ETF_3, the relevant security is a triple-leveraged ETF based on the DAX30. Its price changes, on each trading day, by three times the daily percentage changes of the underlying index DAX30. In treatment ETF_1, in contrast, Security A is simply the DAX30 ETF itself.

The time horizon until maturity of the ETF is 2,000 trading days both for ETF_1 and ETF_3. To generate realized price paths for the two assets, we sample 2,000 consecutive DAX30 closing values, drawn at random from the time period 1964 to 2012.[23] Participants can buy the ETF at a price of 100; if they buy it, they have to hold it for 2,000 trading days. As in Study 2, in both treatments,

[22] Regulatory units and the financial media issued extensive warnings that involve explanations of these counter-intuitive possibilities. The U.S. securities regulator FINRA issued a note in 2009 saying that "[w]hile such products may be useful in some sophisticated trading strategies, they are highly complex financial instruments that are typically designed to achieve their stated objectives on a daily basis. Due to the effects of compounding, their performance over longer periods of time can differ significantly from their stated daily objective..." (Regulatory Note 09-31, Financial Industry Regulatory Authority (2009), page 1).

[23] The instructions in the ETF treatments are analogous to the treatments in Study 2. Additionally, participants receive general information about the DAX30 and a data summary of daily DAX30 movements in the relevant time period. The information is given in the form of a histogram as well as statements specifying the 90% confidence interval ([-1.8% ; 1.8%]) and the overall average of daily percentage changes (0.03%). Note that the participants are UK-based students who typically have little knowledge about German stock markets. Appendix A.1.5 provides an instruction sample.

a participant who chooses Security A receives a fixed bonus if the selling price exceeds a given threshold t_A. These thresholds do not differ between ETF_3 and ETF_1. They are listed in Table 1.4.

	Thresholds for Security A in ETF_3 and ETF_1
Task 1	30
Task 2	60
Task 3	90
Task 4	140
Task 5	200
Task 6	260
Task 7	330
Task 8	450
Task 9	650
Task 10	1,000
Task 11	1,600

Table 1.4: The 11 thresholds.

To elicit three different quantiles, the alternative choice option Security B has three different specifications which are equal to those in Study 2, i.e., Security B1 yields the bonus with 90% probability, B2 with 50% and B3 with 10%. Like in Study 2, each participant faces three choice lists and the choices allow us to infer bounds on the subjectively perceived quantiles.

The computer terminals report feedback to the participants in the form of a sample selling price of Security A. That is, the computer randomly samples a sequence of 2,000 consecutive trading days from the set of all available 2,000-day histories of the DAX30 and uses it to simulate the asset price at maturity. The participants learn the result of the simulation and their payoff. All other aspects of the protocol are identical to Study 2. In both treatments of Study 3, the basic procedure is repeated four times, making for five identical rounds for each participant. 59 participants are in one of the treatments of Study 3 (29 in ETF_3 and 30 in ETF_1), all of them undergraduate students at University College London. The incentivisation structure contains a £5 participation fee and a possible bonus of £5 for each round.

1.5.2 Exponential Growth Bias Prediction

While Study 2 analysed the effects of investment horizon variations, the treatment comparison in Study 3 focuses on the effects of increasing the per-period volatility. We generate the perceived distribution of an EGB decision maker by means of simulations: we randomly sample 500 price paths of ETF_3 and ETF_1 as perceived by an EGB decision maker. By analogy to the previous discussion, the EGB decision maker is assumed to correctly perceive the distribution of daily changes in the relevant asset price, but views them as absolute changes and perceives their distribution to be constant over time. She therefore neglects all compounding. In detail, we simulate the perceived selling price of an ETF by first randomly selecting a start date $t = 0$ at which the price is fixed at $Y_0 = 100$. For each of the ensuing 2,000 trading days $s > t$ we consider the relative change in value on that day, μ_s, and add $(\mu_s - 1)Y_0$ to the current perceived price of the asset: $\tilde{Y}_s = \tilde{Y}_{s-1} + Y_0(\mu_s - 1)$ (with $\tilde{Y}_0 = Y_0$). For example, suppose that $t =$ Jan 2, 1978 was randomly chosen as the starting date and that the DAX30 increased in value by 1.4% on $s =$ Feb 15, 1979. As an EGB decision maker's perceived absolute increase on the latter date, the simulation simply adds $Y_0(\mu_s - 1) = 1.4$ to the price of the asset in the ETF_1 condition (and 3×1.4 in the ETF_3 condition). The simulation thus arrives at a perceived selling price at maturity. Repeating this procedure 500 times for random starting dates generates the perceived distribution of selling prices.

Consistent with the results of previous subsections, the simulations generate the predictions that an increased per-period volatility leads to (i) a larger overestimation of the median return, and (ii) a larger underestimation of the outcome distribution's skew and spread.

1.5.3 Results

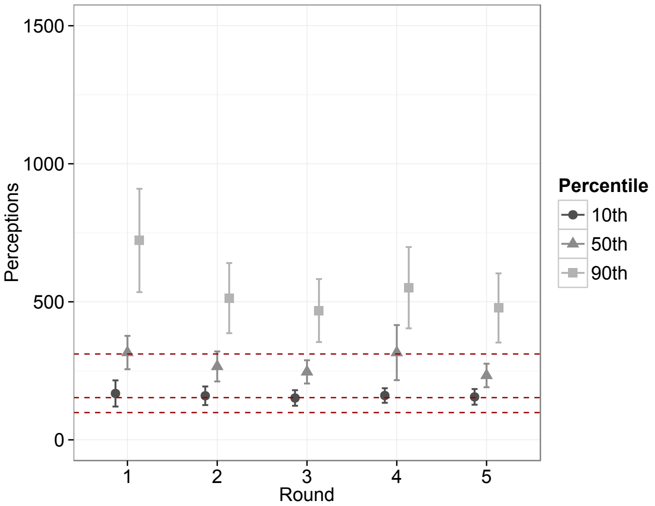


Figure 1.4: Point estimates of the participants' subjective quantiles of Security A's selling price distribution in ETF_1 enclosed by 95% confidence intervals. Dashed lines indicate rational benchmarks for the 10th percentile (lowest), median (middle) and 90th percentile (uppermost).

Figure 1.4 illustrates interval regression estimates of the participants' average perceptions in ETF_1. It shows that for all five rounds average median perceptions (triangle) are significantly above the rational level (middle dashed line). Participants in the experiment are overly optimistic about the simple index ETF. The round-wise arrangement of the three average percentile perception estimates also shows that the perceived distributions are quite symmetric. But we note that with a simple ETF, the true distribution is relatively symmetric as well.

Figure 1.5 captures participants' average perceptions in treatment ETF_3. Again, perceived medians show on average a notable level of overestimation. Moreover, supporting the EGB model, the underlying distributions of subjective percentiles within each round also indicate that the perceived spread and the perceived skewness are on average too small compared to the rational benchmarks. The participants show at least a mild tendency to report skewed distributions but they far underappreciate the actual level of skewness. This misperception is much stronger in ETF_3 than it is in its low-volatility version ETF_1.

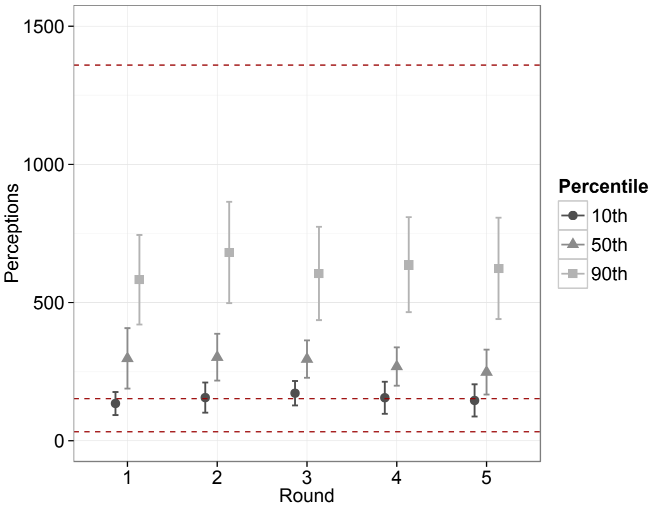


Figure 1.5: Point estimates of the participants' subjective quantiles of Security A's selling price distribution in ETF_3 enclosed by 95% confidence intervals. Dashed lines indicate rational benchmarks for the 10th percentile (lowest), median (middle) and 90th percentile (uppermost).

1.6 Conclusion

This paper investigates how people perceive two important implications of compounding random growth. First, as has been established in the literature, decision makers have a tendency to neglect nonlinear growth. Our experiments add to the evidence of this effect by measuring it in a context with random growth rates. Second, people underestimate the level of asymmetry in growth processes—skewness is "hidden." This is a novel effect in the academic literature (modulo the independent description in Stutzer and Grant (2013)) and may be especially relevant in the context of preferences that consider quantiles of the outcome distribution, like value-at-risk or expected utility with convex utility functions. However, it is important to note that this paper is about misperception, not preferences, and that we measure the effects irrespective of risk attitudes.

Questions about compound interest are, by now, standard procedure in surveys about financial literacy, see e.g., the relevant module in the Health and Retirement Survey documented in Lusardi and Mitchell (2011). The typical evidence is that calculations of multiplicative growth effects show a strong downward bias, often to the extent that all compounding is ignored. One may speculate that if decision makers were trained in the use of log returns, the bias could be reduced. At least some of our experiments give the respondents a very good shot at detecting the nonlinear effects of growth, since we use highly selected and quantitatively skilled students and partially provide them with calculators. It is perhaps all the more notable that we, too, find a strong bias towards linear perceptions.

In Study 3 we also indicate that the effect of skewness is relevant beyond abstract settings. We chose a setting where leveraged ETFs are on offer, as these assets were primarily bought and held by household investors who do not usually have sophisticated risk management tools available. Just as many of these household investors were surprised by the losses they incurred, our participants show a misperception of the effects of leverage, in line with exponential growth bias.

The results may well extend to other contexts of household finance. As we describe in the Introduction, compounding of stochastic growth is often required in contexts of retirement savings. It will be difficult to quantify the effects of the many established anomalies in savings behavior. But progress is made step by step. Our stylized experiments, with full control over information flow and incentives, can at least establish that the expectation of compound distributions deviates predictably from the rational benchmark. Our experiments also show that long investment horizons increase the strength of the misperception.

2

Investment Behavior in Peer-to-Peer Lending

Measuring Applicant Quality to Detect Discrimination

This chapter is based on joint work with Georg Weizsäcker.

2.1 Introduction

Many interesting aspects of gender discrimination concern the interaction of an applicant's gender and quality. Are women's success chances more or less correlated with quality than men's? The answer to this question may well differ from what the unconditional success rates of men and women suggest. Employers, sponsors or other potential discriminators may, for example, favor women over men overall but reward the quality of women's applications less than for men; or they may show the opposite pattern. Allocations and incentives can be severely influenced by such "slope-discriminating" behavior that may appear on top of potential "level-effect" discrimination. Analogous statements are true for analyses that include controls for the quality of applications: the inclusion of interaction terms may bigly change the interpretation of results.

A major difficulty for the analyst is to measure the quality of an application. In most empirical applications, only proxies for applicant quality are available and their interpretation is often all too flexible. The statistical connection between a given proxy and quality may be nonlinear (more generally, difficult to correctly specify) and it may be different for men and women. Both of these effects may also be subject to differential measurement error and to selection effects. A substantial portion of these measurement problems arises simply because the objective of the employer or sponsor is not self-evident. For example, job applicants promise a high-dimensional array of outcomes to the potential employer. It is usually beyond the analyst's power to assess the employers' aggregate valuations of these outcomes.

In this paper, we focus on a narrow financial context, peer-to-peer lending between German households on the online platform *smava.de*. Here, the applicant is a borrower who describes a project and makes a take-it-or-leave-it offer to all potential lenders. The outcome of such a credit application can be reasonably reduced to a single number: its expected internal rate of return. We observe all characteristics of the offered contract and of the applicant that are available to the potential lender, allowing us to assess this measure of predictable quality in detail and with high precision. The nature of the interaction between lender and borrower precludes any other relation between them. Risk considerations are also minimal, due to the platform's specific insurance mechanism, implying that the expected rate of return is a natural candidate for the lender's objective.

Using our inferred measure of quality, we analyze the applicant's chances of success, with a particular focus on the interaction between gender and quality. We address measurement error by modeling the applicant's quality with a detailed structure and by including a statistical correction method (the SIMEX procedure of Stefanski and Cook (1995) adapted to our context). We find significant effects of both slope discrimination and level discrimination. Women have higher success rates than men, conditional on quality, but this gender difference is driven by a larger increase of men's success rate in quality: women appear to get the benefit of the doubt, such that low-quality applications of women are almost equally successful as high-quality applications of women and men. The low-quality applications of men, in contrast, are much less likely to be successful. In terms of chances of a project being fully funded, the success of a below-median-quality application by a man is about half of that of an analogous application by a woman, whereas both genders' above-median-quality applications are equally successful.

Within the larger literature on discrimination, the result is noteworthy in that it confirms a particular feature of (some kinds of) statistical discrimination: if the potential discriminators—here, the predominantly male lenders—find it harder to judge a woman than to judge a man, then weak male applicants should have lower success. This feature is usually not found in discrimination studies, largely because a male-favoring level effect outweighs it. We also show that simpler proxies of quality yield different conclusions. One natural candidate proxy for quality is the applicant's credit rating. Its correlation with success, like that of our own quality measure, suggests that women enjoy positive level discrimination in our context, but not a slope effect. An alternative proxy, the nominal interest rate that the applicant offers, shows an even stronger interaction between gender and quality, such that female applicants do not benefit from higher quality at all. In this sense, one can read our study as cautioning that the choice of proxies for quality is highly important for the conclusion.

A note on causality is in order. We describe correlations, no more. An advantage for simple interpretations of our results is that the available control variables cover essentially all information that is available to the lenders. The lenders therefore plausibly do not condition their choice on other factors. However, we cannot rule out that borrower selection on unobservables (such as a borrower's previous attempts to secure funding) influences the results.

In the next subsection, we briefly review the literature on discrimination with a particular focus on slope discrimination. Subsection 2.3 describes our data context and Subsection 2.4 our econometric modeling. Subsection 2.5 shows the results and Subsection 2.6 concludes briefly.

2.2 Literature Summary and Connections

As indicated above, we distinguish between level discrimination and slope discrimination. In the extant literature, level discrimination is much more frequently investigated, including in studies on differential chances of getting a job offer depending on one's race (Nunley et al., 2014), gender (Kuhn and Shen, 2013), religion (Wright et al., 2013), sexual orientation (Ahmed et al., 2013), age (Riach and Rich, 2010) or attractiveness (Rooth, 2009). The discrimination literature is not limited to hiring decisions, however, as referee decisions in sport (Price and Wolfers, 2010), bidder's choices at eBay (Kricheli-Katz and Regev, 2016), rental market decisions (Carlsson and Eriksson, 2014) and retail market decisions (Zussman, 2013) have all shown to be prone to level discrimination, too. As for the reasons for discrimination, the most prominent proposed theories are Becker's (1957) taste-based or Phelps' (1972) and Arrow's (1973) statistical discrimination (see also Aigner and Cain (1977) for influential early work). Between them, it is notable that taste-based discrimination does more straightforwardly predict level discrimination while the discussion of statistical discrimination much more often includes predictions of slope discrimination. Especially its sub-literature on screening discrimination (Cornell and Welch (1996) and subsequent work) is related to our study. Its main assumption is that the employer can assess certain groups or certain dimensions of applicants better than other groups or dimensions. A low-quality application by a man may, e.g., be less punished for its low quality than in the case of a woman. This specific prediction is, however, essentially unconfirmed in the discrimination literature (see e.g. the explicit statement in Bagues and Perez-Villadoniga (2013)).

More general than in our context, slope discrimination concerns the differential valuation of certain achievements, qualifications or characteristics, evidenced for example in criminal background checks (Pager, 2003) or in the quality of a resumé (Bertrand and Mullainathan, 2004), both of which can be more or less important for different racial groups. Pager (2003) shows that a prior convictions decrease a white job applicant's chances for a call-back by half whereas an African-American ex-convict experiences a reduction in call-back chances by two-thirds, compared to having no criminal record. This is, of course, also in line with the theory of statistical discrimination as the information about a criminal record can be differently informative about the performance on the job of different applicant groups. Such an attribution to a particular mechanism is, however, quite speculative in the absence of a suitable measurement of applicant quality.

Bertrand and Mullainathan (2004) report findings in the opposite direction to that of Pager (2003). They show that having a high-quality resumé is less effective for African-American job applicants than for whites and that the gap between African-American and Whites increases with resumé quality. A possible explanation is that employers may use screening procedures such as not reading further if the job applicant has an African-American name. Employers therefore might not see the correlates of quality of these applicants and thus do not reward them. While Nunley et al. (2014) replicate Bertrand and Mullainathan's (2004) result that differential success chances do not decrease between African-American and white job applicants when additional information about productivity is added to the resumé, Kaas and Manger (2012) find the opposite result, with discrimination of job applicants with Turkish-sounding names in Germany vanishing when a reference letter containing productivity information was added to the application. Again, we remark that all of these results can be explained by statistical discrimination with employers' rationally updating in response to new information. We also remark that the analyzed proxies for quality leave room for interpretation and that one may regard their measurement, and the evidence for their correlations with the employers' objectives, as incomplete.

Discrimination studies that use observational data and estimate output productivity in a more sophisticated way are scarce and usually focus on the labor market. Gallen (2016) uses Danish matched employer-employee data of five industries to estimate the relative productivity of men and women. The gender productivity gap is measured by estimating the efficiency units lost in a firm-level production function for a female worker compared to a male worker, holding other individual characteristics constant. The study finds that 75% of the gender wage gap (-0.16) can be explained by gender productivity differences. Hellerstein et al. (1999) follow a similar methodological path but focus solely on the manufacturing industry. They find a strikingly large gender pay gap of -0.45 but a much more modest productivity gap (-0.16). Azmat and Ferrer (2015) look at productivity data of lawyers. They find that roughly half of the gender earnings gap can be explained by productivity differences. Our paper's main difference to this literature is that we focus on loan application, using observational data. We, too, determine a relatively sophisticated productivity measure, its expected rate of return, and ask whether this measure has a different correlation with success for men and women. This arguably reduces the room for interpretation significantly, even relative to the conventional productivity measures in the labor market literature.

Prior research on peer-to-peer lending markets includes studies that focus on the borrowers' financial strength (such as credit grade), paid interest, or loan duration as covariates of funding success (Iyer et al. (2009), Herzenstein et al. (2011), Avery et al. (2004)) as well as studies that focus on personal characteristics such as race, age, gender, or beauty (Theseira (2008), Pope and Snydor (2011), Ravina (2012)). Gender is usually just used as a control variable in these studies. In contrast, Barasinska and Schäfer (2014), which is also the peer-to-peer lending study most similar to our paper, specifically look for gender discrimination in the same data of the platform *smava.de*. Their regression results show no significant relationship between an applicant's gender and the likelihood of being fully funded. While our approach differs from theirs in multiple ways, the fact that we find very different results despite using the same data is another indication that the method of measuring quality is key in analyses of discrimination.[24]

2.3 The Peer-to-Peer Lending Process

2.3.1 Participation and Information Conditions

The data set consists of loan applications posted between March 2007 and March 2010 on the German peer-to-peer lending platform *smava.de*.[25] In contrast to traditional bank lending, peer-to-

[24] Other studies using the online lending platform *smava.de* as data source are Barasinska (2011), Kraus (2013) and Pötzsch and Böhme (2010).

[25] The data are suitable for our purposes only until March 2010, whereafter a loan application can involve joint liability of two persons living in the same household, with only aggregated financial information being shown on the platform.

peer lending refers to direct lending between private persons. The platform sets the rules according to which the lending process is carried out and supplies the infrastructure of the web-based intermediation procedure. Although funds are exchanged (almost) directly between persons no further communication between peers is possible. Borrowing and lending is carried out anonymously. Informational asymmetries are high since investors have only access to a limited set of verified information provided by the platform.

In order to participate in the lending process, be it as investor or as borrower, adult members of Germany's general population can register at *smava.de* and verify their identity via the "postident" procedure administered by the German postal service provider. *smava.de* requests personal information such as one's name, gender, age and place of residence, of which only age and gender are published on the platform in an uncensored way. Borrowers and lenders choose a user-name and the user's address is opacitized in the sense that only the federal state of residence is published. The platform also collects financial information of potential borrowers to infer their "Schufa rating" and "KDF indicator", both being published on the platform. The Schufa rating is the main German credit rating for individuals, issued by the German national credit bureau. It reflects a person's default risk on a scale from A (lowest risk) to M (highest risk). The KDF indicator measures a person's financial burden from outstanding consumer loans ranging between 1 (lowest financial burden) and 4 (highest financial burden).[26]

The available verified information is complemented by voluntary provided unverified information. That is, on the platform one can choose one of several preselected options indicating one's employment status and, after a successful registration, one can voluntarily upload a profile picture and write a text describing oneself in greater detail.

Registrants at *smava.de* have to be older than 18 years and must permanently reside in Germany. A person can register only either as investor or as borrower, not as both. For eligibility to borrow, one must prove a monthly income of €1,000 at least, a Schufa rating below *Category I* and a debt-to-disposable income ratio no higher than 67%.

2.3.2 The Terms Of An Investment

A successfully registered borrower can post a loan application that describes the requested loan amount (in multiples of €500 with a maximum of €50,000), the term of the loan (either 36 or 60 months), the annual interest rate he or she is willing to pay (in multiples of 0.1 percentage points), the purpose of the loan (an unverified choice from a fixed list of 17 categories) and a voluntary added text providing further (unverified) details about the loan application.

Based on the observable loan and borrower specific information a potential investor can decide whether or not to invest in a given loan application. A possible investment has to be in multiples of €250 and must not exceed the requested total loan amount or €25,000.[27] In contrast to other peer-to-peer lending platforms such as Prosper.com where loans are auctioned, *smava.de* works on a first-come-first-served basis. As soon as the requested loan amount equals the aggregated supply of funding, or after 14 days, the loan application is closed. Investors can not underbid each other. However, loan applicants can raise the offered interest rate during the bidding period to attract more funding. The final rate is the effective rate for all investors.

If less than 25% of the requested amount are funded after the loan is closed, the application is withdrawn from the platform and the received bids (if any) are transferred back to the investors. In this case, the potential borrower can change the terms of the loan application and post a modified

[26] The specific calculations used by *smava.de* are as follows. The sum of all monthly payments to loans (including loans from *smava.de*) is divided by the monthly disposable income (without savings), resulting in the debt-to-disposable income ratio. This measure is assigned to categories from 1 to 4, constituting the KDF indicator. The relevant information as well as the indicator are continuously updated. Debt payments, income, and savings are not observable for other peers apart from the summary in the KDF indicator.

[27] Most investors provide only a small fraction of the requested loan amount so that loans are usually financed by many different investors.

application.

If the supplied loan amount covers at least 25% of the requested sum, investors are contractually bound to realize their bids accordingly. Upon acceptance of the funds by the loan applicant the loan contract is legally valid. After a legal loan contract is achieved, *smava.de* charges investors and the borrower a fee.[28]

2.3.3 The Risk Of An Investment

After a loan is paid out, the borrower is contractually bound to repay the funds in constant monthly annuities. A borrower can repay early but, in this case, has to compensate the investors for the missed interest payments. If a borrower fails to repay, investors incur a loss.[29] That is, loans arranged by *smava.de* are not secured by collateral or third parties. But two insurance mechanisms apply. First, the claim to the resulting debt from a default is sold to a collecting agency which, if successful, recovers an average share of roughly 20% of the missed payments for the investors. Second, and more importantly, a risk sharing mechanism is installed, functioning as follows. Risk sharing is effective via loan pools. Investors are pooled by two characteristics of the loans they have invested in: the loans' Schufa ratings and their durations until maturity. For example, all investors who put their money in loan applications with a repayment duration of 60 months and a Schufa rating "B" are assigned to one pool. With two different loan durations and eight applicable Schufa ratings, the grouping results in a total of 16 different pools. The monthly principal repayments are aggregated within each of the pools and each investor receives a share proportional to his or her relative investment in the respective pool. If a loan defaults, the resulting loss is subtracted from the pooled repayments. Thus, each member of the pool partially compensates for the loss. The resulting repayment rates per pool range from 99% for pools of loans with Schufa rating A to 84% for those with a Schufa rating H.[30] Interest payments are exempt from the pooling procedure and transferred directly to the investors if the loan did not default in the previous month.

Overall, the risk sharing mechanism essentially insures against the loss of principal, at the cost of the effective pay-out lying below the nominal interest rate. As we show in the next section, more than 80% of completed contracts show repayments that corresponds to an internal rate of return lying within the interval [0.05, 0.1] and the remaining 20% all lie within the interval [-0.05, 0.2]. We regard the distribution of repayments as small enough to justify an analysis under the assumption of investor risk neutrality: an investor's main challenge is to identify the probability of default, which would induce a loss of interest payments from the time of default onwards. This assessment mirrors expected-utility calculations under standard assumptions, showing that risk considerations can be neglected.[31]

[28] The structure of the fees for borrowers and investors has been changed by *smava.de* once during the period that is relevant for our study. Prior to February 2009, there was no fee for investors. Borrowers paid simply 1% of the borrowed amount. Afterwards, investors were charged €4 per bid and borrowers 2% of the borrowed amount (or at least €40) for loans with a 36 month duration and 2.5% (or at least €40) for loans with 60 months duration. We account for this change in our loan return calculations and in our regressions via fixed effects.

[29] A failure of repayment is declared as soon as the monthly payment is 60 days late.

[30] The data represent historical average payment rates over the period from April 2007 to January 2010, published by *smava.de*.

[31] For example, consider as a benchmark an investor who evaluates a safe 1-period investment of €8,000 (the empirical average of requested loans in our data set) that repays at an interest of 6.6% (the empirical average of realized internal rate of return). To investigate the potential importance of risk aversion, consider increasing the pay-out variance to the *maximal* extent that is possible such that (i) the expected IRR is held constant and (ii) the possible return realizations are within the range observed in our data set. Under a log utility evaluation, such an increase in risk affects the investor's expected utility by less than half of what a one-percentage-point reduction in the safe interest rate (down to 5.6%) would inflict. In other words, even an unrealistically high degree of risk would have minuscule effects on investors.

2.3.4 The Data

Our data set consists of 4144 closed loan applications submitted by 3400 borrowers, including the online-published personal and financial information of the borrowers.
On average 73% of the loan applicants are male, with one half living in the northern states of Germany. Average borrower age is 44 and the mean requested sum is roughly €8,000, offered at an average nominal 9.9% annual interest. The distribution over the 4 KDF-indicator categories is 1:17%, 2:24%, 3:33%, 4:26%. The distribution of individual borrowers' Schufa ratings is A:15%, B:16%, C:9%, D:10%, E:11%, F:12%, G:16%, H:11%.
At the end of our observational period, 5671 lenders were registered on the platform and submitted on average 10 funding decisions. Only 625 (11%) of lenders are female, precluding us from performing a meaningful interaction of our analysis with lender gender. The exact descriptives are shown in Table 2.1.

2.4 Empirical Strategy

In the following, we analyze whether and how lenders evaluate loan applications differently for male versus female applicants. For exposition, we start by presenting two null hypotheses that describe investor behavior as aiming at maximal expected returns, conditional on the variables that are observable to him or her. The main idea behind these hypotheses is that if we were to include a perfect measurement of an application's quality as a control, then investor behavior should not show any partial effects of applicant gender. (While our measurement is not perfect, we argue that it is unbiased and that the measurement error is small and correctable, see Subsection 2.4.3.) Deviations from these nulls allow multiple interpretations, including the possibility of taste-based discrimination or differential consideration of information about male versus female applicants. But they rule out statistical discrimination in combination with rational expectations.[32]
In specific, we ask: "Is there a significant correlation between the received funding and a loan applicant's gender?" (Hypothesis 1, on level discrimination) and "Does the correlation between expected return and received funding differ by the applicant's gender?" (Hypothesis 2, on slope discrimination).
To measure a loan application's quality, we use the expected internal rate of return conditional on the available information, E(IRR). We describe this measure in detail in Subsection 2.4.2. The dependent variable of the analysis is a loan application's received funding share, categorized in the three relevant categories that the platform imposes, {category 1 - funding <25%, category 2 - funding ≥25% and <100%, category 3 - funding =100%}.[33] We fit their incidence in an ordered logit framework (see Subsection 2.4.1). Using the β_X coefficient estimators of explanatory variables X in these regressions, we formulate our hypotheses:

Hypothesis 1: *A loan applicant's gender is not related to funding success, i.e.,* $\beta_{Gender} = 0$.

Hypothesis 2: *A loan application's interaction of expected return with the applicant's gender is not related to funding success, i.e.,* $\beta_{Gender*E(IRR)} = 0$.

[32] Under the assumption that lenders aim at maximal *subjectively* expected returns, the deviations must stem from false subjective expectations that may depend on the applicant's gender. See Weizsäcker (2010) for analogous tests of rational expectations for different types of information sets in social learning experiments.

[33] We have more fine-grained information on the extent of funding but since almost 90% of the loan applications received either 100% of the requested amount or nothing at all, a finer analysis makes little difference.

2.4.1 The Ordered Logit Approach

In our econometric specification we model each loan application i as being of perceived utility U_i to the aggregate pool of investors at *smava.de*. We assume that the observable attributes of a loan application determine U_i according to

$$U_i = \beta^{'} x_i + \varepsilon_i \tag{2.1}$$

where x_i depicts loan application i's observable attributes, $\beta^{'}$ describes the weights (or coefficients, which can be negative) of these attributes and ε_i is a random disturbance term (with a standard logistic distribution in our application).

To capture equation 2.1 econometrically, we make use of the ordered logit framework. Since U_i is not directly observable, we collapse its range of possible values into categories. That is, we define a variable $Y_i \in \{1, 2, 3\}$ which measures the funded share of a loan application in three observable categories, as defined above. The categorization depends on whether U_i passes a threshold κ_m, with $m = 1, 2$. The parameter vector β is then to be estimated together with the thresholds, κ_m, via maximum likelihood with

$$P(Y_i{>}m) = \frac{exp(X_i\beta - \kappa_m)}{1 + exp(X_i\beta - \kappa_m)}. \tag{2.2}$$

2.4.2 Expected Return Calculation

A loan's expected internal rate of return (irr) is unobservable for us, as well as for investors, and we therefore use the available loan and borrower specific information to come up with an estimate.[34]

For each loan there exist $T + 1$ different outcome scenarios, with T as the total number of monthly payments during the repayment period. The possible scenarios range from an immediate default before the first payment is due to a complete repayment as planned.[35]

Each of the possible outcome scenarios occurs with a certain probability. The probability that the borrower defaults in the first month is denoted $p_1 = Pr\{D = 1\}$, with D as a discrete random variable indicating the month of default, and analogously for defaults in other months, up to the best case scenario being $p_{T+1} = 1 - Pr\{D \leq T\}$. These probabilities are unknown but can be estimated based on past defaults of borrowers in our data set. We estimate the default risks of all months, $p_1...p_{T+1}$ with a discrete time hazard model, using as explanatory variables the observable characteristics of the borrower and the terms of the loan. With the help of this information, for each loan we calculate how likely each possible default scenario occurs.[36]

Next, we use the contractual repayment structure of *smava.de* to determine the monetary values of each possible outcome scenario. All loans are annuity loans, i.e., repayments consist of a principal repayment part and an interest payment part. Due to the collective insurance mechanism implemented by *smava.de* (described in Subsection 2.3.3), investors receive an insured part of the principal repayments no matter whether the borrower defaults or not.[37]

The insurance mechanism does, as described above, not include the interest payment part. Overall,

[34] For similar estimation attempts in a peer-to-peer lending context see Barasinska (2011) and Kraus (2013).

[35] It is not possible to leave out a certain monthly repayment and continue repaying later on. If a borrower defaults once, it's over.

[36] For further details see Appendix A.2.1.

[37] As described above, the degree of insurance depends on the investment pool that the investor belongs to. At the time of investment investors do not know the precise default rate in their pool, mainly as it may vary over time. *smava.de* publishes a continuously updated estimate of the pool-specific default rate and in our calculation we make the simplifying assumption that the ensured part is known to be constant at this level.

repayment is

$$\text{Payoff}(D) = Rate_{pool} \times \sum_{t=1}^{T} Principal_t + \sum_{t=1}^{D-1} Interest_t, \tag{2.3}$$

with $Rate_{pool}$ as the insured fraction of the annuity repayments and the outcome scenario indicated by D, with $D = 1, ..., T+1$. Using the payoff we can solve for the irr for each of the possible outcome scenarios. It is given by solving the condition for a break-even flow of payments,

$$\text{Investment} + \text{Fee} = \sum_{t=1}^{T} \frac{Annuity_t}{(1+irr)^t}, \tag{2.4}$$

where Investment is the amount invested, Fee is the amount charged for using the platform and $Annuity_t$ is the sum of the partly insured principal payment ($Rate_{pool} \times Principal_t$) and the uninsured interest payment ($Interest_t$) received in period t.
The *expected* internal rate of return to an investment, $E(irr)$, is the sum of the possible returns irr_t that would arise from the possible $T+1$ outcome scenarios weighted by the respective default probabilities, $p_1, ..., p_{T+1}$:

$$E(irr) = \sum_{t=1}^{T+1} p_t \times irr_t, \tag{2.5}$$

with E(IRR) $= 12 \times E(irr)$ as the annual return measure. Figure 2.1 illustrates the resulting density of E(IRR) over the posted loan applications in our data set.

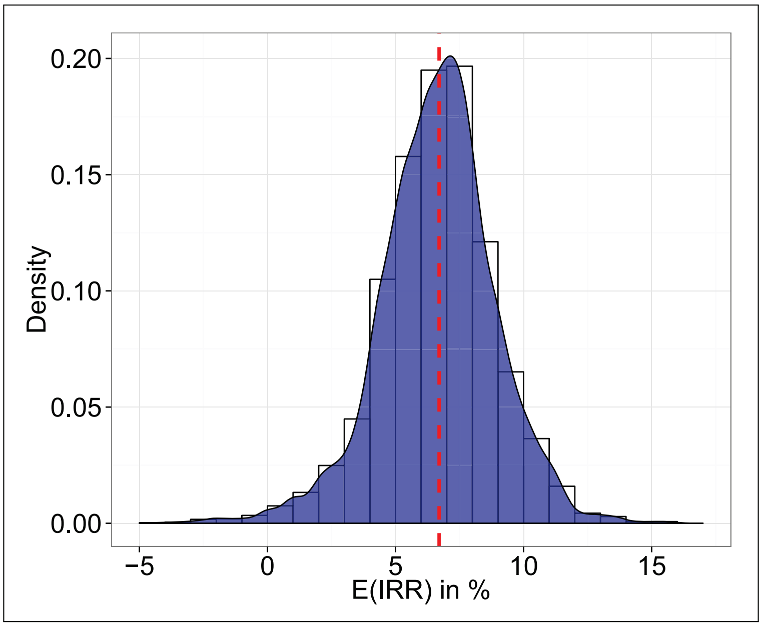


Figure 2.1: Density of E(IRR) over the posted loan applications in our data set, with the horizontal dashed line indicating the mean return at 6.6%.

2.4.3 Correcting for Measurement Error

Incorporating likely quantities of not directly observable variables in an econometric analysis comprises the risk of biased results, an unreliable coverage level of confidence intervals and a reduction of statistical power due to measurement error. To account for measurement error, the literature proposes different approaches such as the semi-parametric correction technique by Sepanski et al. (1994), the two-stage bootstrap method by Haukka (1995), the regression splines approach of Berry et al. (2000), the regression calibration method by Hardin et al. (2003), the so-called indirect method by Jiang and Turnbull (2004) and the adjusted estimator of Cameron and Trivedi (2005), among others.

The method most appropriate for our setting—due to its robustness to distributional assumptions, its applicability to relatively small samples and its performance in several simulation studies (see, e.g., Fung and Krewski (1999))—is the *simulation-extrapolation method* (SIMEX) developed by Cook and Stefanski (1994). It allows retrieving asymptotically unbiased and efficient estimates for regression based models. The method requires, importantly, an estimate of the measurement error variance (or, as in our case, of the estimation error variance), σ_u^2. Since we construct the quality measure ourselves, we know the measurement process which is (potentially) subject to estimation error and can use a Monte-Carlo approach to arrive at the estimate.[38]

Equipped with an estimate of the measurement error variance, the basic idea of the SIMEX method is fairly straightforward: if a causal relationship is biased by measurement error, then adding more measurement error should increase the degree of this bias. That is, the measurement error variance σ_u^2 is increased to $(1+\Gamma)\sigma_u^2$ where Γ controls the amount of added measurement error. By adding successive levels of measurement error, it is possible to estimate the relationship between the expected value of the coefficient and Γ, and then extrapolate back to the unbiased estimate. That is, for different values of $\Gamma = (0.5, 1, 1.5, 2)$ we create $b = 1, ..., B$ pseudo data sets via simulations

$$\mathrm{E}(\tilde{\mathrm{IRR}})_{b,i} = \mathrm{E}(\tilde{\mathrm{IRR}})_i + \sqrt{\Gamma}\mathrm{Normal}(0, \sigma_u^2)_{b,i} \tag{2.6}$$

where $\mathrm{E}(\tilde{\mathrm{IRR}}) = \mathrm{E}(\mathrm{IRR}) + u$ and $u \sim N(0, \sigma_u^2)$, with E(IRR) as the true value. Then we refit the pseudo data to obtain the bth pseudo estimate

$$\hat{\theta}_b(\Gamma) = \hat{\theta}(\{Y_i, \mathrm{E}(\tilde{\mathrm{IRR}})_{b,i}\}_1^n) \tag{2.7}$$

and take the average over all B simulations

$$\hat{\theta}(\Gamma) = B^{-1}\sum_{b=1}^{B}\hat{\theta}_b(\Gamma). \tag{2.8}$$

We then write the estimates as a function of Γ and since $\Gamma = -1$ is the case of no measurement error, extrapolating back to $\Gamma = -1$ gives the parameter estimates for this case.[39] The employed extrapolation function is a quadratic polynomial which is usually the default setting in most statistical packages and has a good performance in many cases.

2.5 Results

Towards answering Hypotheses 1 and 2, we first give a short data overview and then estimate the above-described ordinal logit model and apply the SIMEX correction.[40] Finally, we compare our

[38] We re-estimate our measure with different parameter specifications which we randomly sample from their confidence intervals that are estimated from our data set (using the procedure described in Appendix A.2.1). This accounts for the uncertainty in the employed parameter estimates, which might result in measurement error. The resulting data allows us to estimate the variance parameter(s) of the estimation error of our quality measure. We estimate the mean estimation error variance for the full sample and individually for each loan application. The reported results below use the former, full-sample estimator to run the SIMEX analysis. None of the conclusions change qualitatively if we use the individual-loan estimators.

[39] Although the SIMEX approach with its simulation character seems a natural fit for the bootstrap to obtain standard errors of the SIMEX-parameter estimates, for complex models like ours it is simply not feasible. We instead use the jackknife method developed by Stefanski and Cook (1995) which has a much smaller computational burden and has shown to deliver valid estimates in simulations.

[40] The ordinal logit model is estimated with the *polr* procedure using the *mass* package in R and implemented in a modified version of the *simex* 1.5 R package (see, Lawrence (2009)). More information concerning the

choice of explanatory variables with natural other candidates.

2.5.1 Data Overview

Female loan applicants get on average 2.7% more of their requested loan amount funded then males (Two-sample t test, p-value<0.001). There are, however, many other characteristics of a loan application besides applicant gender that could explain this finding. Table 2.1 gives a descriptive overview of the main variables that we observe, separated by gender. It shows that female loan applicants offer on average higher interest rates, request lower amounts, have higher Schufa ratings and are older than male applicants.

Loan/Borrower Characteristics	Female	Male
E(IRR) in %	6.70	6.60
Offered Loan Rate in %	10.14***	9.78***
Loan Duration in months	50.34	49.92
Requested Loan amount in €1,000	7.47***	8.17***
Schufa Rating in Scores from 1-8	4.51**	4.36**
KDF Rating in Scores from 1-4	2.69	2.69
Borrower Age in Years	47.01***	43.21***
Borrower Residence in North Germany in %	50.01	49.30
Length of Project Description in Letters	395.04**	371.57**
No. of Observations	**1114**	**3030**
Lender Characteristics	Female	Male
Lender Age in Years	45.33***	41.49***
Lender Residence in North Germany in %	30.08*	32.69*
No. of Observations	**625**	**5046**

Significance levels: $*$: <10% $**$: < 5% $***$: < 1%

Table 2.1: Two-sample t test results for average gender differences of borrowers and lenders. P-values refer to one-sided tests.

2.5.2 Econometric Analysis

We specify three different econometric models and estimate each of them with and without measurement error correction by means of SIMEX. Model (1) simply controls for a loan applicant's gender with the dummy variable $borrower_{male}$ valued 1 if the respective person is male and zero otherwise. Model (2) additionally accounts for the loan application's expected return by means of the mean centered variable E(IRR) and its interaction effect with the gender variable ($borrower_{male}*$ E(IRR)). Model (3) adds further controls and their respective interactions with E(IRR): the applicant's age $borrower_{age}$ (which is a mean centered continuous variable) and place of residence $borrower_{north}$ (a dummy variable valued 1 if the respective person lives in the northern part of Germany, and zero otherwise).[41] Moreover, Model (3) includes controls for all other variables that we were able to cleanly extract from out data set: the requested loan amount, the loan rate, the (categorical) purpose of the loan, the loan applicant's Schufa rating, the loan applicant's KDF indicator, the loan applicant's occupation, the length of the loan description and *smava.de*'s fee structure. Table 2.2 list the results.

implementation in R available upon request.

[41] We define the two geographical areas such that the number of inhabitants and the economic output are approximately equal in both clusters. Federal states assigned to the northern part of Germany are Bremen, Hamburg, Berlin, Schleswig Holstein, Mecklenburg Western Pomerania, Saxony-Anhalt, Brandenburg, Lower Saxony, and North Rhine-Westphalia while the remaining federal states form the southern part.

Dependent Variable: $share\_funded$			
Explanatory Variable	(1)	(2)	(3)
$borrower_{male}$	-0.203**	-0.134	-0.166
	(0.09)	(0.09)	(0.11)
E(IRR)		0.049	0.169*
		(0.03)	(0.09)
$borrower_{male}$* E(IRR)		0.185***	0.110**
		(0.04)	(0.04)
$borrower_{age}$			-0.015***
			(0.01)
$borrower_{age}$* E(IRR)			-0.001
			(0.00)
$borrower_{north}$			0.116
			(0.09)
$borrower_{north}$* E(IRR)			-0.033
			(0.04)
Fixed Effects	No	No	Yes
	SIMEX-Model(1)	SIMEX-Model(2)	SIMEX-Model(3)
$borrower_{male}$	-0.203**	-0.018	-0.192*
	(0.09)	(0.09)	(0.10)
E(IRR)		0.104*	0.503***
		(0.05)	(0.14)
$borrower_{male}$* E(IRR)		0.502***	0.318***
		(0.06)	(0.06)
$borrower_{age}$			-0.013***
			(0.00)
$borrower_{age}$* E(IRR)			-0.006***
			(0.00)
$borrower_{north}$			0.112
			(0.09)
$borrower_{north}$* E(IRR)			-0.098
			(0.05)
Fixed Effects	No	No	Yes

Significance levels: $*$: <10% $**$: $<$ 5% $***$: $<$ 1%

Table 2.2: Ordinal logit results with (below) and without (above) correction for measurement error via the SIMEX method. Fixed effects and κ coefficients are reported in Table A.14 in the appendix.

Hypothesis 1 - level effect: All model specifications exhibit a negative coefficient for a loan applicant being male. However, only for the two Model (1) versions is this correlation strongly significant while for Model (2) and Model (3), at most a marginal significance is detected. In Model (1) it could obviously be the case that gender picks up the influence of other variables, like higher E(IRR), consistent with the gender differences reported in Table 2.1. Controlling for E(IRR) in the saturated Model (2) and adding various other controls in Model (3) does not change the the sign of the gender coefficient but reduces its significance. Overall, the results of model estimations with and without measurement error correction do not allow us to fully reject the null of Hypothesis 1 but at least cast serious doubt at its validity. Women appear to be at an advantage overall, in our data set.

Hypothesis 2 - slope effect: The evidence of a significant interaction between quality and gender is much stronger. For all of our model specifications does the coefficient for the interaction term between gender and E(IRR) show a positive sign at a strong significance both in the case of no measurement error correction and after a SIMEX application. The SIMEX correction increases the size of main effects and interaction effects of E(IRR).[42]

[42] This is not surprising since the SIMEX approach eliminates some of the noise due to measurement error for the E(IRR) variable and thus decreases its bias towards zero.

For Model(2), the gender difference in the importance of E(IRR) is so strong that the coefficient of women's E(IRR) is only marginally significant and this significance vanishes if measurement error controls are omitted. In other words, the application's predictable quality is a much better indicator of the funding probability for men than for women. Thus, the regression results allow us to reject Hypothesis 2. A natural interpretation is that the effect reflects irrational expectations, as indicated in the literature of statistical discrimination and screening discrimination: The predominantly male lenders have a worse understanding of women than of men and therefore can identify low-quality men but not low-quality women.

Our data allow us further to check for slope effects of two other borrower characteristics with a discrimination potential, i.e., a borrower's age and place of residence. While the expected return is not evaluated differently for borrowers living in the north or south of Germany a borrower's age seems to matter. The coefficient for the interaction with the expected return in SIMEX-Model (3) shows a significant negative effect. The size of the predicted age effect is, however, clearly smaller than the gender difference even if a change of one standard deviation for age (13.6 years) is considered. Further analogous estimations of SIMEX-Model (2) for the two variables confirm these findings (see Table A.13 in Appendix A.2.2).

Using the predicted values from SIMEX-Model (2), Figure 2.2 graphically summarizes our findings regarding gender discrimination at *smava.de.* For loan applications in the +/- 4-percentage-point range around the average of expected returns, the likelihood of funding success and the E(IRR) measure correlate much stronger if the application was submitted by a man rather than by a woman. This difference vanishes for loan applications with expected returns exceeding the sample average by more than 4%points. However, the relevance of this group for our analysis is limited since it accounts for only 3% of our data set.

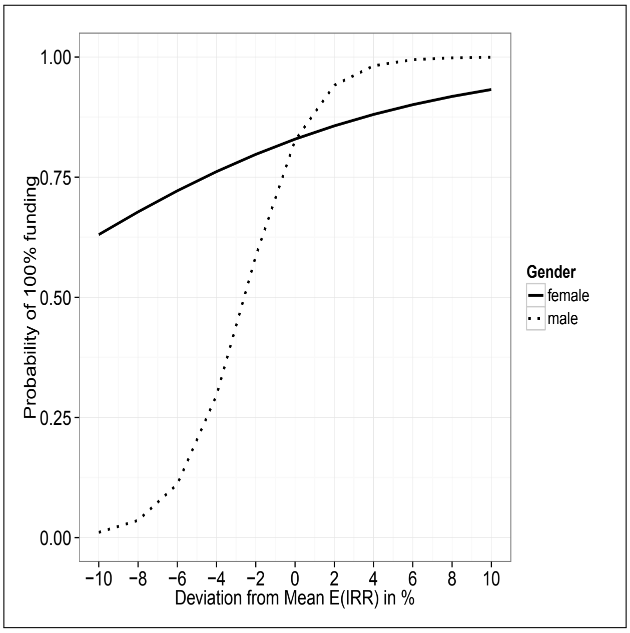


Figure 2.2: Predicted values for gender differences in the probability of full funding over expected return deviations from the mean.

2.5.3 Comparison of Candidate Proxies for Quality

We close our empirical analysis by comparing our measure of quality, E(IRR), with other possible proxies of an application's quality. Would our conclusions change if one uses different proxies to measure applicant quality? The following figures indicate that the answer is affirmative.

As a benchmark, Figure 2.3 depicts the predicted values of Model (2), i.e. using the same explanatory variables as in Figure 2.2 but without the SIMEX correction (coefficients: $E(IRR)$ (0.049), $borrower_{male}$ (-0.134), interaction term (0.185***)).[43] Comparisons of the two figures shows that both reveal the same qualitative insights but that the positive slope for men's E(IRR) is stronger with the SIMEX correction.

In contrast, for Figure 2.4, the role of E(IRR) is taken by another natural candidate for a quality proxy, the offered loan rate that is contained in the description of an offer. With this measure of quality, one would arrive at the conclusion that females are, if anything, harmed by offering higher quality. The gender difference in the relation between quality and funding success would thus be even bigger than when using E(IRR) as a quality measure (coefficients: $loan\_rate$ (-0.002), $borrower_{male}$ (-0.505*), interaction term (0.030)).

Figure 2.5 uses yet another measure of quality, the Schufa credit rating. Here, we detect no slope difference across genders at all, only a level effect appears (coefficients: $Schufa$ (-0.116**), $borrower_{male}$ (-0.141), interaction term (-0.016)).

In sum, the comparison of the different quality measures serves as a robustness check but also as a warning: while the rough pattern of results may be similar, the interpretation may change significantly depending on the choice of quality controls. Additionally, the significance of the gender level effect depends on the model specification. While the $borrower_{male}$ coefficient is negative for all three specifications only for the offered loan rate measure does it show to be significant. This makes it even more pressing to measure the quality or value of an application in the best possible way, and if possible account for the precision of the measurement.

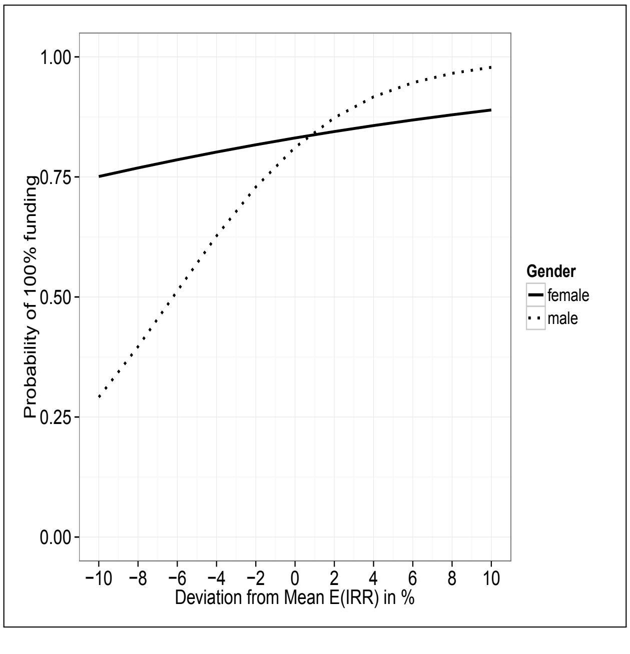


Figure 2.3: Predicted values for gender differences in the probability of full funding over expected return deviations from the mean (without SIMEX correction).

[43] For the other two measures of quality that we study, the SIMEX correction is unavailable and we therefore take Figure 2.3 as a benchmark.

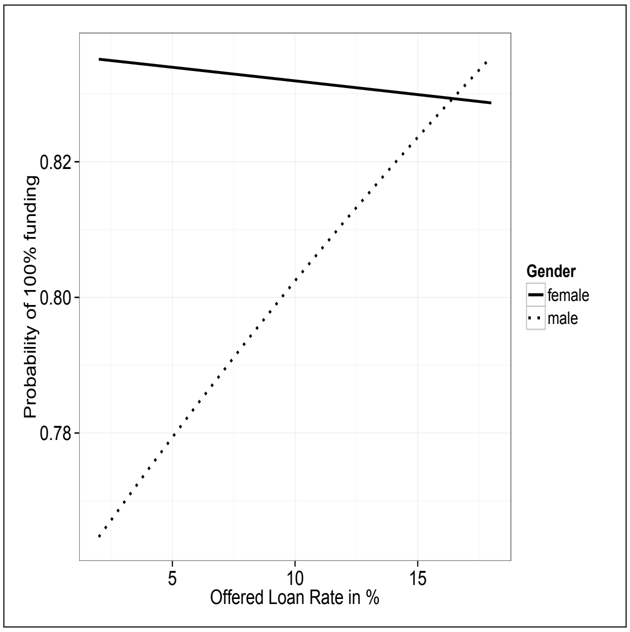


Figure 2.4: Predicted values for gender differences in the probability of full funding over offered loan rate.

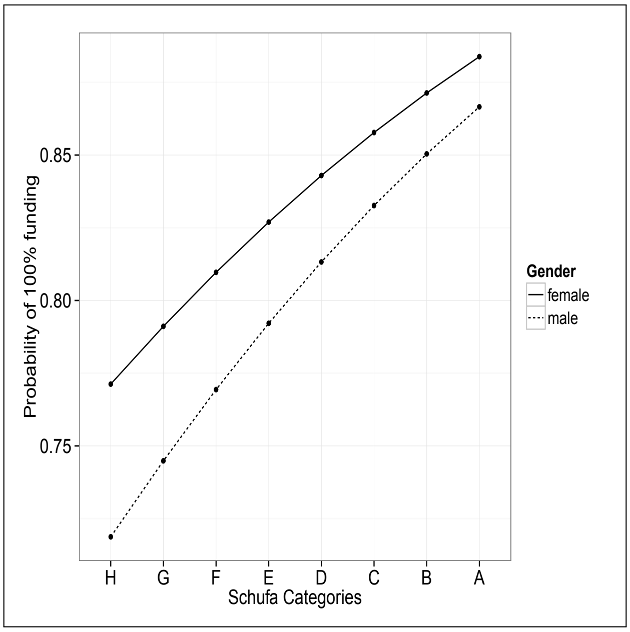


Figure 2.5: Predicted values for gender differences in the probability of full funding over Schufa rating.

2.6 Conclusion

The paper makes mainly methodological contributions in that it illustrates in novel ways that the choice of a proxy for applicant quality, and its measurement, play a role in the conclusions that one may draw from empirical analyses of discrimination contexts. But of course, the substance of the analysis may be the main interest to most readers. The finding that low-quality applications of men are penalized more than those of women has immediate analogues to many other contexts, and is in line with the straightforward predictions of statistical discrimination: men can judge men better than they can judge women and this is bad for weak male applicants.

Our choice of data context, peer-to-peer lending, has specific and noteworthy characteristics. First, the “rules of the game” are transparent, not only to the agents but also to the analyst. Especially the fact that the information conditions are highly controlled is a key advantage in order to detect gender differences while controlling for key information on the applications' qualities. Second, the context may well be different from labor market contexts in terms of the prevailing stereotypes and tastes regarding the two genders. Although it is hard to compare across settings, one may find it plausible that an investor who uses peer-to-peer lending has a much more positive view of female loan applicants, compared to an employer's view of female job applicants. We agree to this and clearly do not claim to have found a general women-favoring pattern in discrimination behavior.

3

Binary Choice Belief Elicitation

An Adaptively Optimal Experimental Design

3.1 Introduction

For choice under uncertainty the true probabilities of the alternatives are seldom available in real-life. Usually, people choose based on subjective beliefs. According to economic theory, agents should form such beliefs by following the laws of probability and make decisions based on the expected utility of the alternatives weighted by their subjective beliefs (Savage, 1954).

How subjective beliefs are actually formed and related to choices under uncertainty, however, remains one of the most important and challenging questions not only for economic phenomena, such as deviations from Nash equilibrium strategies (Costa-Gomes and Weizsäcker, 2008) or precautionary savings (Guiso et al., 1992), but for almost every area of everyday life, such as attending preventive medical checkups (Shiloh et al., 1997) or vaccinating children (Mills et al., 2005).

To allow researchers to investigate the link between people's beliefs and their choices, most large-scale surveys, such as the CentER Pension Barometer and the Health and Retirement Survey (HRS), elicit subjective beliefs from respondents. In a survey environment, however, the feasibility of the elicitation tasks and the quality of the data depends crucially on the comprehensibility of the tasks by the survey respondents, the time needed to conduct the tasks, and the accuracy with which the elicitation design can elicit beliefs.

A very important practical problem in the elicitation of subjective beliefs is that survey respondents cannot be relied upon to provide answers that satisfy the laws of probability. Some of the most common mistakes are documented by Delavande and Rohwedder (2008), van Santen et al. (2012), and Binswanger and Salm (2013) who find that a substantial share of respondents' probability statements violate additivity (i.e., the sum of the probabilities of two disjoint events is not equal to the probability of the union of the events) and/or monotonicity (i.e., an event has a higher probability than another event of which it is a subset). Kezdi and Willis (2008) even find that many HRS respondents state different probabilities for the same question if asked twice within 20 minutes.

Obviously, people have great difficulties with reporting probabilities, either in the form of densities or cumulative probabilities. This causes problems in later analyses. Inconsistent answers are difficult to deal with econometrically and such data is often simply thrown out of the sample. Excluding inconsistent answers may, however, lead to selection bias as violations are usually not random (van Santen et al., 2012). Other studies report far lower violation rates, but this is mainly because the question design either makes it harder or even impossible to provide inconsistent answers (see, e.g., Dominitz and Manski (1997), Delavande and Rohwedder (2008), and Huck et al. (2014)). Forcing respondents to be consistent by design makes all data usable and thus may help avoid the selection problem, but it may also introduce substantial measurement error and, again, bias estimates.[44]

In this paper, I present a novel econometric approach for eliciting beliefs that confronts respondents with nothing more than binary choice questions. I ask experimental participants only which of two events they judge to be more likely. Rather than assessing the (cumulative) probabilities of events (Dominitz and Manski, 1997) or assigning events to given probabilities (as in the bisection method of Abdellaoui et al. (2011)) respondents only have to choose one of two events.

Research in psychometrics suggests that binary judgments have desirable properties in experiments. Schneider et al. (1974), for example, experimentally show that subjects' binary comparisons of the loudness of sounds satisfy transitivity and monotonicity. Further support comes from the observation that people are better at evaluating probability intervals than generating them (Winman et al., 2004), which is consistent with the finding that relative judgments are easier made than absolute ones (Gigerenzer and Selten, 2001). Therefore, I expect that – especially for general population samples, where the average respondent is unlikely to have much if any mathematical or statistical training – answering binary choice questions may be a much easier task than stating

[44] Quantifying this measurement error and taking explicit account of it is becoming an increasingly important consideration in empirical applications (see, e.g., Drerup et al. (2014)).

subjective probabilities or even constructing confidence intervals.

My methodology is very general and can be applied to a variety of domains (e.g., environmental variables like the weather or market prices like stock market returns) as well as probabilistic structures of almost arbitrary complexity. It can be used to estimate the relative likelihoods of discrete events or to estimate the entire probability distribution over a continuous state space as, indeed, is the case in this paper's application.

A stochastic choice model assumes that respondents have a subjective probability distribution from which they generate noisy binary judgments. Given enough binary comparisons and assumptions on the functional form of the probability distribution and the distribution of the noise term, the parameters of the underlying distribution can be estimated. The econometric method is agnostic about the size of the systematic component of responses and can accommodate and, in fact, identify anything from responses that are strictly consistent with the laws of probability to responses that are made entirely at random.

The psychological simplicity of the binary question design does, however, come at a cost: binary choices, by their very nature, contain very little information. While existing methods elicit probabilities and can therefore cardinally order events, binary events allow only for a (noisy) ordinal ordering. Therefore, I use the stochastic choice model not only to estimate the parameters of the distribution *ex post* but also to choose the experimental questions *ex ante*. I choose questions such that their answers contain maximal information in a well-defined statistical sense and do this adaptively: After any history of responses the model is estimated and the next question is chosen optimally *given the data*.

Model simulations show that despite the low informational content of binary choices, this adaptive design allows to recover the parameters of the choice model extremely well, even with comparatively few judgments. Furthermore, the results of an online study demonstrate the user-friendliness of the design, subject pool densities of the belief parameter estimates with reasonable values, and correlations of the elicited beliefs and other related measures that imply a satisfactory goodness of elicitation.

The rest of this paper is organized as follows. Subsection 3.2 briefly discusses related methodologies. Subsection 3.3 introduces the econometric framework. Subsection 3.4 describes the implementation of the adaptively optimal questionnaire design, Subsection 3.5 reports the simulation study results, Subsection 3.6 covers the online experiment, and Subsection 3.7 concludes.

3.2 Related Methodologies

Numerous methods for the elicitation of subjective beliefs can be found in the literature. These methods differ from my approach in several respects, including the sort of responses elicited (measures of central tendency or probabilities in contrast to binary judgments), the way the responses are processed (constructing point estimates or confidence intervals in contrast to the estimation of complete probability distributions), the number of questions asked, whether participants are restricted in their judgments, and whether every respondent is asked the same questions.

Dominitz and Manski (1997), for example, ask respondents about the probabilities with which they think certain events will occur. Chaining multiple of such questions in principle allows one to construct an entire probability distribution.[45] However, plenty of evidence documents that in answers to such "chained" questions a substantial share of responses are inconsistent with the laws of probability (see, e.g., Binswanger and Salm (2013), Delavande and Rohwedder (2008), van Santen et al. (2012), Kezdi and Willis (2008), Gouret and Hollard (2011), Weinstein and Diefenbach (1997), Riddel and Shaw (2006), and Jakus et al. (2009)).

[45] For similar elicitation applications see, e.g., Hurd et al. (2011), Hurd and Rohwedder (2012), and Arrondel et al. (2014).

An easy fix to avoid inconsistent responses (e.g., violations of additivity or monotonicity) is to simply ask only a single question. In the HRS, however, Binswanger and Salm (2013) find that many of such single probability statements bunch at subjective probabilities of 0%, 50%, or 100% raising questions about whether respondents meant to express these beliefs exactly, whether they meant to express beliefs "in the ballpark" or whether these responses simply reflect an inability or an unwillingness to give an answer at all.

Mechanisms used in the experimental economics literature to encourage honesty and a careful assessment of the question by the respondent are so-called *scoring rules* (Garthwaite et al., 2005). They enable an evaluation of a probability statement by assigning a (numerical) score based on the stated subjective probability of an event and the actual realization of that event. Simply put, scoring rules can be thought of as a reward that the respondent wishes to maximize.[46] Over time, quite a large literature has evolved, with scoring rules building on different assumptions about the respondent and the nature of the assessed event.[47] In a survey environment, however, a scoring rule application is typically not feasible since the elicited probability statements are not immediately verifiable. Additionally, in contrast to the typical (student) subject pool in economic lab experiments, most survey respondents might have problems to correctly apprehend the incentive scheme behind a certain scoring rule (see, e.g., Jenkinson (2005) and Delavande et al. (2011)).

Delavande and Rohwedder (2008) propose an alternative elicitation method that avoids inconsistent answers by design. An entire subjective distribution over future social security benefits is elicited using a visual computer-based tool. Respondents are shown an interface with seven bins into which they are asked to distribute 20 balls so as to represent their probabilistic beliefs. This makes giving inconsistent answers impossible and makes the whole data set usable. It may, however, also introduce measurement error: respondents who do not understand the question or do not have well-defined probabilistic beliefs will be forced to supply an answer that is consistent with the laws of probability and be indistinguishable in the data from those who have well-defined probabilistic beliefs. The authors find that one fifth of participants assigns all probability mass to the bin that contains their (previously stated) point forecast and that 73% of participants use only one or two bins.[48] The authors also unfold bins in up to five finer grained bins if respondents use only one or two of the original bins resulting in 80% of those presented with two unfolding bins allocating all mass into one bin and two-thirds of those presented with three, four, or five unfolding bins using just two of them. While the authors interpret this behavior as a willingness to give more precise estimates when asked, it could also reflect people's unwillingness to give any sort of probabilistic answers at all.

Rothschild (2011) refines the design of Delavande and Rohwedder (2008) by dynamically changing the binning: If a respondent places more than 50% mass in a bin this bin is split in two so as to allow the respondent to express more fine-grained beliefs. However, this condition for a split may not be the most compelling: If the bin is in the center of the subjective distribution, then exactly how the mass is distributed within the bin will not matter very much if one is interested, e.g., in the dispersion of the distribution. In these cases, splitting a bin in the tail of the distribution may result in a larger informational gain even if that bin only contained much less probability mass.[49]

Another branch of methods attempts to overcome the limitations of the above described "direct" methods by indirectly eliciting subjective probabilities from choices involving lotteries or gambles. One of the most popular "indirect" approaches is the method of "external reference events," which confronts subjects with a list of lottery pairs (see, e.g., Merkhofer (1982), Offerman et al. (2009), Andersen et al. (2014), and Menapace et al. (2015), among others). Each pair consists of a lottery

[46] In a Bayesian context, such scores are often described as utilities referring to the Bayesian principle of maximizing the expected utility of a probability statement (see, Chapter 2 in Bernardo and Smith (1994)).

[47] For a discussion of the most prominent scoring rules and their assumptions see Schlag et al. (2015) and for a technical review with a larger focus on statistical properties see Gneiting and Raftery (2007).

[48] This is not due to the bin containing the point forecast being the middle bin. In a separate study of Delavande and Rohwedder (2008) the third (rather than the central fourth) bin was the one that contained the point forecast and there were no significant differences in the distributions elicited.

[49] For a review of further applications and modifications of this belief elicitation approach see Goldstein and Rothschild (2014).

that is characterized by the event whose subjective probability needs to be elicited and another lottery that is represented by an external reference event whose objective probability is disclosed to the subject. Indifference between two lotteries is an indication that both events are assigned the same probability by the subject (Spetzler and Staël Von Holstein, 1975).

However, this elicitation technique might be prone to bias since subjects have to simultaneously assess two different sources of uncertainty. For example, Kilka and Weber (2001) and Abdellaoui et al. (2011) experimentally show that subjects' choices depend on the source of uncertainty with which they are confronted. The more diverse the involved sources of uncertainty the complexer the choice environment for the subjects resulting in a possibly biased subjective probability elicitation (Baillon, 2008).

The class of the "exchangeability methods" attempts to overcome source dependence and was first described by Raiffa (1968). Rather than being confronted with probabilities, subjects just have to deal with magnitudes of the event. These methods are based on Ramsey's (1931) and de Finetti's (1937) basic idea of exchangeable events, i.e., any given respondent should be equally willing to bet on either of two exchangeable events — roughly, two events she holds to be equally likely to occur. To elicit a subjective probability distribution, the state space is split in two events and the respondent is asked to choose between bets on these two. Following the choice, another twofold partition is generated from the chosen event and again bets are offered. This is repeated until the respondent becomes indifferent between the two events.

This paper is closely linked to Baillon (2008), the first experimental study that uses the exchangeability approach to non-parametrically estimate subjective probability distributions.[50] Most similar to my experimental design, however, is a modified version of the exchangeability approach introduced by Welsh et al. (2008). The authors use repeated relative judgments (along with confidence statements) to non-parametrically construct subjective probability distributions. Respondents have to indicate which of two randomly chosen values is closer to their subjective estimate. Additionally, they are asked to state their confidence about their choice ranging from 50% (guessing) to 100% (certain). A subjective distribution is constructed by distributing probability mass over value ranges based on respondents' relative judgments and weighted by the confidence statements. This is repeated 10 times each time with randomly chosen values from the range of values not ruled out by previous answers. Following the last question, the final distribution is corrected in order to satisfy the definition of a probability density function.

The key difference between the current literature of the exchangeability approach and my proposed experimental design is that not only do I use a stochastic choice model to estimate mean and standard deviation parameters of the individual subjective outcome distribution while allowing respondents to answer the experimental questions with imperfect accuracy, but I also maximize the informativeness of the questions, thereby designing an optimal questionnaire (individually for each participant) and minimizing subjects' effort.

3.3 Econometric Framework

3.3.1 Stochastic Responses

As probability judgments tend to be noisy I specify a stochastic choice model that assumes that people have some distribution in mind from which they generate noisy responses. This subjective distribution is approximated to be a $\mathcal{N}(\mu, \sigma^2)$. The subjective probability for any interval event A is

[50] For further applications of the the exchangeability approach see, e.g., Abdellaoui et al. (2011), Cerroni et al. (2012), and Cerroni and Shaw (2012).

then $P(A|\mu, \sigma) = \Phi(A_u|\mu, \sigma) - \Phi(A_l|\mu, \sigma)$, with $\Phi(\,\cdot\,)$ the normal cumulative distribution function and A_u and A_l the upper and lower bounds of the interval.

I assume, furthermore, that a respondent faced with a comparison between event A and event B will *not* always choose the event she holds to be more likely but that this decision will be subject to error (see, e.g., Luce (1958) and Laskey and Fischer (1987)). To formalize this, I employ a standard Fechner model, as common in the empirical literature on stochastic choice (see, e.g., Hey and Orme (1994)). Formally, the respondent bases her choice not on $P(A|\mu, \sigma) - P(B|\mu, \sigma)$ but on $P(A|\mu, \sigma) - P(B|\mu, \sigma) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \eta^2)$ is a random error term. She will choose event A if $P(A|\mu, \sigma) - P(B|\mu, \sigma) + \varepsilon > 0$ and choose event B if $P(A|\mu, \sigma) - P(B|\mu, \sigma) + \varepsilon \leq 0$. This can be interpreted as an ordinary response error. The log-likelihood of a single observation q for an individual i choosing between event A and event B then is

$$\begin{aligned} LL_{iq}(\mu_i, \sigma_i, \eta_i) = {} & 1_{\{choice_q = A\}} \\ & \cdot \ln(\Phi(P(A|\mu_i, \sigma_i) - P(B|\mu_i, \sigma_i)|0, \eta_i)) \\ & + 1_{\{choice_q = B\}} \\ & \cdot \ln(1 - \Phi(P(A|\mu_i, \sigma_i) - P(B|\mu_i, \sigma_i)|0, \eta_i)). \end{aligned} \tag{3.1}$$

For $q = 1, ..., Q$ binary judgments per individual this results in $\sum_{q=1}^{Q} LL_{iq}(\mu_i, \sigma_i, \eta_i)$ as log-likelihood for individual i's response data.

3.3.2 Bayesian Estimation Framework

I embed the above stochastic choice model in a Bayesian framework with prior distributions specified for the three parameters μ, σ, and η. These priors are updated after each binary judgment according to Bayes' rule. This serves as the basis for my estimation strategy.[51]

My experimental data have a panel structure with a relatively large number of units of observation (respondents) and relatively few or noisy data from each of these units (binary judgments). The estimation of any model of choice using such data confronts a fundamental trade-off. Pooling noisy responses and estimating the parameters of the model as if multiple respondents were one results in more precise estimates. Such complete pooling is common, e.g., in the estimation of risk preferences (see e.g., Holt and Laury (2002)). However, complete pooling excludes, by assumption, the possibility of individual heterogeneity in parameters, thus, potentially leading to bias (Rossi and Allenby, 2003). Alternatively, parameters can be estimated for each decision maker separately and independently. If they are based on only a few observations these estimates are bound to be imprecise.

As a middle ground between these two extremes estimates can be *partially* pooled. One way of achieving such partial pooling is by means of a hierarchical model in which the parameters of each unit of observations have some freedom to vary but in which estimates are shrunk toward the average parameters in the population. By following this approach, hierarchical Bayes allows for a simultaneous estimation using the complete subject pool without treating multiple respondents as one. The parameters governing the prior distributions for each of the respondents are neither fixed nor identical, but are themselves drawn from distributions. The priors on the model parameters are referred to as first-stage priors and the priors on the "hyperparameters" (i.e., the first-stage prior parameters) as second-stage priors. In preference measurement designs, such a simultaneous estimation has improved estimation accuracy in cases of few observation per experimental subject (Allenby and Rossi, 1998).[52]

I use the hierarchical estimation approach below only to overcome small sample issues related

[51] For a review on experimental designs employing a Bayesian framework, see Chaloner and Verdinelli (1995).

[52] For a more detailed description of the hierarchical Bayes approach and additional applications see Chapter 5 in both Gelman et al. (2005) and Rossi et al. (2006).

to the inherent difficulty of estimating the probability of what are essentially tail events with very few observations. This is a major issue when estimating the error parameter η form my experimental data. However, for the estimations used in the process of finding, first, the optimal next question and, subsequently, the subjective individual μ and σ parameters, I employ a separate estimation approach. This is because (thanks to my optimal elicitation approach) the elicited data are informative enough to allow me to control for individual heterogeneity.

3.3.2.1 Separate Estimation Approach

As a benchmark model, I estimate the parameters separately for each respondent and for each domain. For the priors of respondent i's three parameters μ_i, σ_i, and η_i, I use fixed values as prior parameters (marked by ˜) denoted as $\tilde{\mu}_0$, $\tilde{\sigma}_0$, and $\tilde{\eta}_0$.[53] The joint posterior distribution of the three parameters for respondent i is then given by Bayes' rule as:

$$\begin{aligned} P(\mu_i, \sigma_i, \eta_i | \text{data}, \, \tilde{\mu}_0, \, \tilde{\sigma}_0, \, \tilde{\eta}_0) \quad \propto \quad & P(\text{data}|\mu_i, \sigma_i, \eta_i) \\ & \cdot P(\mu_i| \, \tilde{\mu}_0) \\ & \cdot P(\sigma_i| \, \tilde{\sigma}_0) \\ & \cdot P(\eta_i| \, \tilde{\eta}_0), \end{aligned} \tag{3.2}$$

where $P(\text{data}|\mu_i, \sigma_i, \eta_i)$ is the likelihood of the response data of individual i and $P(\mu_i| \, \tilde{\mu}_0)$, $P(\sigma_i| \, \tilde{\sigma}_0)$ and $P(\eta_i| \, \tilde{\eta}_0)$ are the priors for μ, σ and η, respectively.[54]
I use Markov Chain Monte Carlo (MCMC) methods to sample from the posterior distribution and draw inferences from this sample.[55]

3.3.2.2 Hierarchical Estimation Approach

As an alternative to the separate estimation strategy, I employ a hierarchical Bayes approach as introduced in Subsection 3.3.2.[56] After pooling responses and additionally (compared to the benchmark model in Subsection 3.3.2.1) specifying prior distributions on the hyperparameters (i.e., on the first-stage prior parameters $\tilde{\mu}_0$, $\tilde{\sigma}_0$, and $\tilde{\eta}_0$), Bayes' rule again applies and gives the joint posterior distribution of the three sets of individual parameters $\{\mu_i\}$, $\{\sigma_i\}$, and $\{\eta_i\}$ for all

[53] For reasons of a simpler notation, only a single prior parameter is considered here. In my application below, however, this is changed to two, with a normal prior for μ being governed by $\mathcal{N}(\tilde{\mu}_\mu, \, \tilde{\sigma}^2_\mu)$, and σ and η following an inverse gamma distribution with $\mathcal{IG}(\tilde{\alpha}_\sigma, \, \tilde{\beta}_\sigma)$ and $\mathcal{IG}(\tilde{\alpha}_\eta, \, \tilde{\beta}_\eta)$, respectively. How exactly prior parameters are set is described below together with the estimation results.

[54] Note that for an uninformative uniform prior on all three parameters, the joint posterior would simply be the likelihood function and the estimates identical to those in a frequentist framework.

[55] MCMC methods are necessary because the posterior distribution given prior and likelihood is not available in closed form.

[56] It is expected to improve the estimation accuracy for the error parameter η since the probabilities of tail events (such as choosing the less likely interval under the subjective distribution) are usually difficult to estimate from low informative data without pooling observations.

individuals $i = 1, ..., I$ and the population estimates for the hyperparameters:

$$\begin{aligned} P(\{\mu_i\}, \{\sigma_i\}, \{\eta_i\}, \tilde{\mu}_0, \tilde{\sigma}_0, \tilde{\eta}_0|\text{data}) \quad \propto \quad & P(\text{data}|\{\mu_i\}, \{\sigma_i\}, \{\eta_i\}) \\ & \cdot P(\{\mu_i\}|\, \tilde{\mu}_0) \\ & \cdot P(\{\sigma_i\}|\, \tilde{\sigma}_0) \\ & \cdot P(\{\eta_i\}|\, \tilde{\eta}_0) \\ & \cdot P(\tilde{\mu}_0) \\ & \cdot P(\tilde{\sigma}_0) \\ & \cdot P(\tilde{\eta}_0), \end{aligned} \tag{3.3}$$

with $P(\text{data}|\{\mu_i\}, \{\sigma_i\}, \{\eta_i\})$ as likelihood for the response data of all I individuals; $P(\{\mu_i\}|\, \tilde{\mu}_0)$, $P(\{\sigma_i\}|\, \tilde{\sigma}_0)$, and $P(\{\eta_i\}|\, \tilde{\eta}_0)$ as first-stage priors; and $P(\tilde{\mu}_0)$, $P(\tilde{\sigma}_0)$, and $P(\tilde{\eta}_0)$ as second-stage priors. In order to fully exploit the hierarchical structure of the model and to let the values of the hyperparameters be mainly determined by the data, the second-stage priors are specified to be as uninformative as possible.[57] Posterior distributions are, again, found via MCMC methods.

3.4 Adaptively Optimizing Questionnaire Design

I use my stochastic choice model embedded in a Bayesian framework not only to estimate the parameters of the model after the conclusion of the experiment but also to estimate the model *during* the experiment to maximize the information content of each question asked. The central idea behind the dynamic questionnaire design is, given a (possibly empty) set of previous responses, to find the parameter estimates and, based on these estimates, compute for each possible next question some measure of informational content. The question maximizing this measure is then picked by the dynamic design to be the next question asked.

I follow the approach of Toubia et al. (2013), which is derived from dynamic design methods of the literature on preference measurement (see, e.g., Abernethy et al. (2008), Toubia et al. (2003), and Sawtooth Software (1996)) and maximize the expected determinant of the parameter Hessian.[58] [59] This ensures that the asymptotic covariance matrix of the parameters has a minimal determinant and, thus, the uncertainty concerning the parameters is as low as possible.[60]

3.4.1 Optimal Question Sequences

To be more specific, assume a respondent i has already answered q questions. My task is now to find the optimal pair of events $(A_{i(q+1)}, B_{i(q+1)})$ for question $q + 1$ according to the measure of informational content. Hence, I have to evaluate every possible pair of events and choose the

[57] This is achieved by employing uniform distributions for the second-stage priors, with ranges covering all reasonably possible values. First-stage priors are defined as for the separate estimation approach described in Footnote 53.

[58] A similar measure of informational content is the Kullback-Leibler divergence used by Wang et al. (2010). However, Wang et al.'s (2010) approach requires a discretization of the parameter distribution. In contrast to that, following Toubia et al. (2013) enables me to deal with continuous parameter spaces.

[59] Different norms of the Hessian can be maximized, such as the absolute value of the largest eigenvalue, the trace norm, etc. I chose this measure of informational content based on its computational feasibility for my binary approach and its performance in related tasks (see, Toubia et al. (2013)).

[60] Under general conditions, the asymptotic covariance matrix of the maximum likelihood estimator (MLE) is equal to the inverse of the Hessian of the log-likelihood function at the MLE (see, e.g., McFadden (1974) and Newey and McFadden (1994)).

one that maximizes the expected determinant of the Hessian of the posterior after $q+1$ questions (Toubia et al., 2013):

$$p_A \cdot |\det(H_{iq} + h(A^{A=1}_{i(q+1)}, B^{B=0}_{i(q+1)})| + p_B \cdot |\det(H_{iq} + h(B^{B=1}_{i(q+1)}, A^{A=0}_{i(q+1)})|, \tag{3.4}$$

where

$$p_A = \Phi(P(A|\mu_{iq}, \sigma_{iq}) - P(B|\mu_{iq}, \sigma_{iq})|0, \eta_{iq}) \tag{3.5}$$

is the probability of respondent i choosing event A based on the posterior after q questions, and $H_{iq} + h(A^{A=1}_{i(q+1)}, B^{B=0}_{i(q+1)})$ is the Hessian of respondent i's posterior after $q+1$ questions if event A is chosen.

To approximate equation 3.4, I estimate the choice probabilities and the Hessians based on *maximum a posteriori* **(MAP)** parameter estimates (i.e., the modal values of the posterior distribution) (De Groot, 1970) and not based on the entire posterior distribution after q questions. While integrating over the entire distribution would obviously be the correct thing to do, this is computationally infeasible. Thus, after q questions have been answered by an individual, i, the computer program has to (a) update the MAP estimates of respondent i and (b) compute for all possible next questions the expected value of the determinant of the Hessian evaluated at the MAP estimates and select the one question maximizing this criterion. This has to be done for all $i = 1, ..., I$ respondents after each of the $q = 1, ..., Q$ questions.

In order to reduce waiting times in the online experiment (and also for the simulation study), I calculated all possible questionnaire trajectories *ex ante*. For 20 questions the resulting question-tree branches out with $2^{19} - 1 = 524,287$ knots and at each knot steps (a) and (b) have to be performed. Having done this previously, I just have to read from my *ex ante* calculations which next question follows which history.[61]

3.5 Simulation Study

Before I collect data in an online experiment to test my procedure with real people, I first check how well my approach recovers true parameters under ideal circumstances in a series of simulations. That is, the simulations test parameter recovery of subjective beliefs for simulated experimental participants. I vary the levels of response error, the number of questions asked, and the difference between prior and true parameters. I am interested in the rate of convergence and the accuracy of the estimates after a maximum of 20 questions. In an additional set of simulations, I compare my optimal adaptive experimental design to a simple random selection of questions.

Simulation results show that the adaptive procedure is extremely effective at increasing the informational content of the questions asked. In many cases significantly fewer than 20 questions are necessary to estimate the parameters of the model to high precision.

[61] The software used to compute these steps was Python 3.4 and in specific its package PyMC 2. A description how to analytically derive the Hessians to make use of PyMC in an efficient way is outlined in Appendix A.3.2.

3.5.1 The DJIA Domain

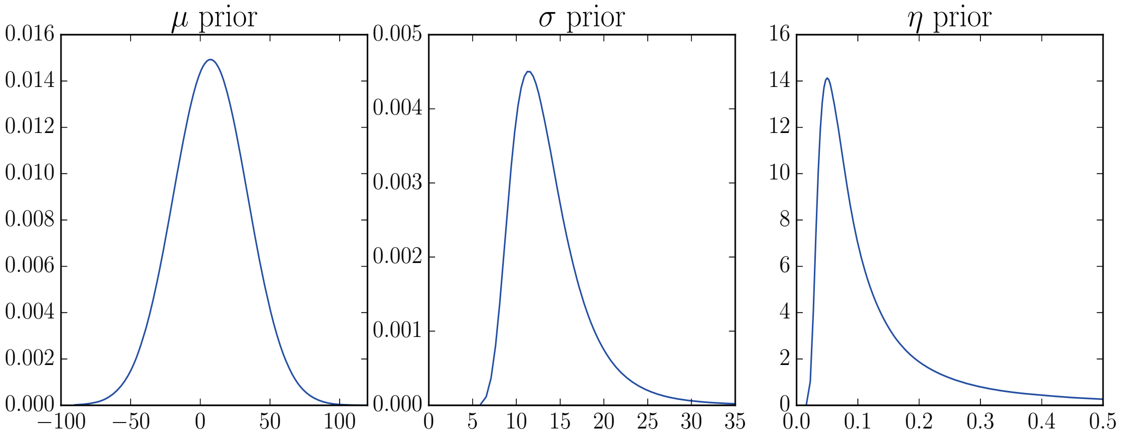


Figure 3.1: Prior distributions for μ, σ and η in the DJIA domain.

In all simulations I have simulated respondents make judgments about the year-on-year percentage returns of the Dow Jones Industrial Average (DJIA). The binary judgments concern intervals from an overall grid ranging between [−70% ; +120%], with a step size of 5 between [−40% ; +60%] and a step size of 10 between [−70% ; −40%] and [+60% ; +120%]. The optimal binary judgments are picked from the set of all possible pairs of non-overlapping intervals that have endpoints on this grid, resulting in 31,465 event pairs.[62]

The prior densities used as described in Subsection 3.3.2.1 are illustrated in Figure 3.1. I chose a normal prior for the μ parameter and inverse gamma priors for the other parameters σ and η. The normal prior is centered on the historical mean (7.3%) with a standard deviation such that the distribution's probability mass encloses 95% of the range of 250 trading day returns of the DJIA over the last 50 years. The prior distribution for σ has its mode at the historical DJIA standard deviation (11.4%). For η I set the mode at 0.05, i.e., people with a response-error rate at the modal value get the difference of probabilities wrong by at most 5% with 65% probability and get it wrong by at most 10% with 95% probability. While these priors are informative, they are not very strong, as is shown later.

3.5.2 Individual Level Simulation

To get an in-depth look at the mechanics of my optimal adaptive experimental design, Figure 3.2 shows questions and answers along an optimal sequence of 20 binary judgments for a simulated respondent. It also illustrates how the estimated subjective belief distribution about the future DJIA evolution changes following each judgment. The upper panel shows the sequence of pairs of intervals that were presented to the simulated respondent. The intervals are represented as bars that show both the location and the size of the intervals (e.g., the first pair of intervals is [−70% ; +5%] and [+10% ; +80%]). A bar is colored in green if the simulated respondent chose

[62] The different step sizes over the whole interval serve the purpose of decreasing the number of non-overlapping intervals below 40,000 while still covering the historical range of 250 trading day returns of the DJIA over the last 50 years. This reduction is necessary since for each of these event pairs I have to compute the expected determinant of the Hessian at each of the $2^{19} - 1 = 524,287$ knots of the question-tree. Even though this looks similar to a standard *embarrassingly parallel problem* in parallel computing since the problem can only be separated into a limited number of parallel tasks at each knot, I have to make this minor restriction.

the corresponding interval correctly (i.e., the chosen interval is the one with the higher probability under the subjective distribution) and in red if the respondent chose the interval in error (the response error being governed by $\eta = 0.05$).[63]

The lower panel shows the estimated underlying distribution as it evolves along the sequence of choices. The estimated distribution starts out at the prior modal values with $\mu = 7.3$ and $\sigma = 11.4$, here shown in yellow, and then moves in successively darker blues toward the true distribution, which is further to the right and much more spread out with $\mu = 30$ and $\sigma = 20$ (shown in green). As the posterior estimates move to the right so do the pairs of intervals that the respondent is asked about. The final estimate of the subjective outcome distribution, after only 20 binary choice questions, is strikingly close to the true one.

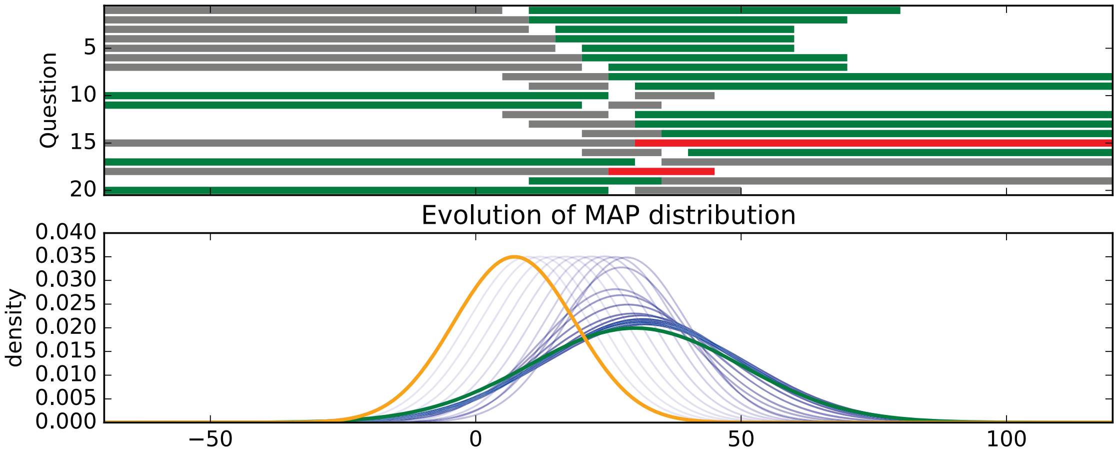


Figure 3.2: Simulation results for a single simulated respondent.

3.5.3 Simulation Study Design

Simulations are performed on a $2 \times 2 \times 2$ grid: They examine parameter recovery for all possible parameter combinations of 2 different values for $\mu \in \{10, 30\}$, $\sigma \in \{12, 20\}$ and $\eta \in \{0.05, 0.15\}$. For each combination 500 sequences of questions are simulated, each containing 20 binary judgments. The simulated binary judgments are made according to the decision rule described in Subsection 3.3.1. After each judgment, q, is made by a simulated respondent, this respondent's parameter vector is estimated based on all her judgments by finding the MAP parameter estimates (marked by $\hat{}$ below). The next optimal question $q + 1$ is then chosen as described in Subsection 3.4.

[63] The binary judgments picked by the adaptive design become increasingly difficult to answer over the course of the 20 questions resulting in more frequent errors at later questions.

3.5.4 Simulation Results

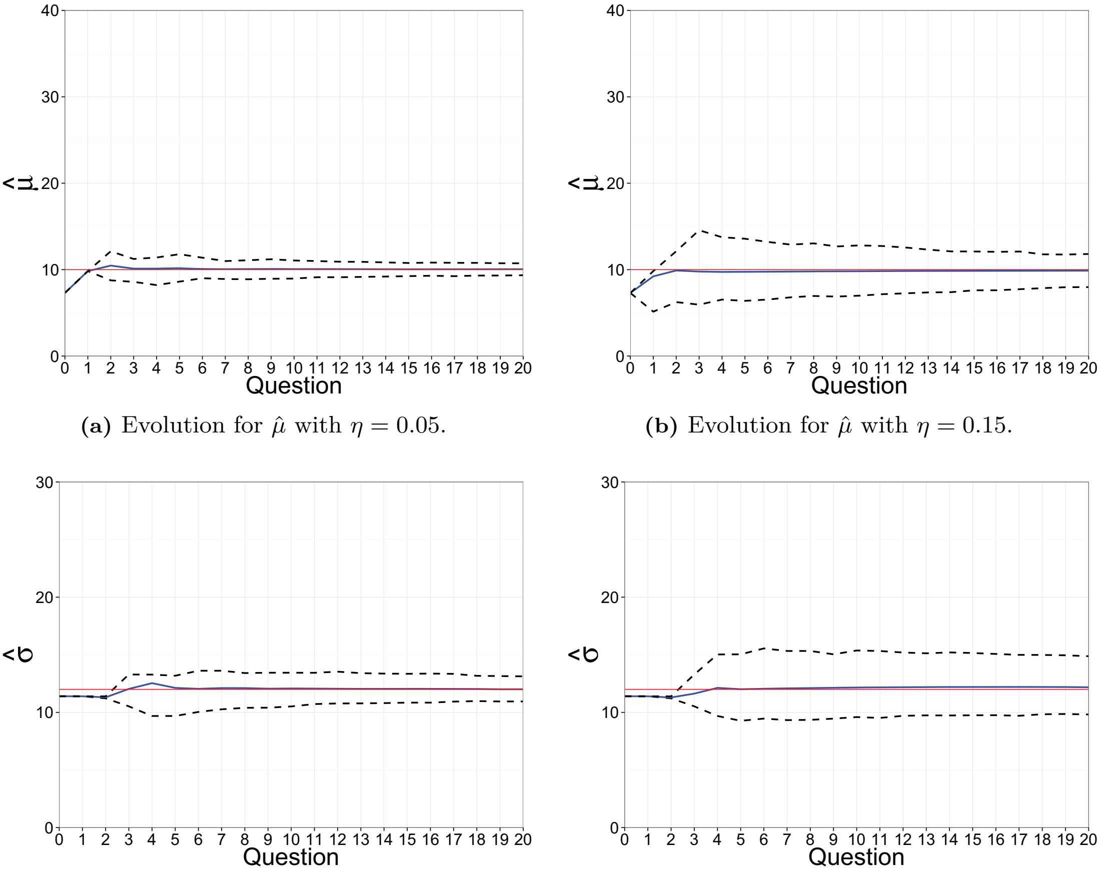
Parameter specification			Number of questions asked				
μ	σ	η	0	5	10	15	20
10	12	0.05	0.160	0.059	0.048	0.039	0.033
10	20	0.05	0.350	0.171	0.066	0.052	0.042
30	12	0.05	0.403	0.202	0.043	0.040	0.029
30	20	0.05	0.593	0.391	0.238	0.080	0.041
10	12	0.15	0.160	0.125	0.105	0.093	0.083
10	20	0.15	0.350	0.248	0.177	0.145	0.121
30	12	0.15	0.403	0.202	0.052	0.074	0.063
30	20	0.15	0.593	0.392	0.245	0.147	0.111

Table 3.1: Average absolute percentage error between simulated MAP estimates and true values for optimally chosen questions.

For each of the 2 × 2 × 2 parameter combinations and each of the 500 simulated sequences of questions I calculate the absolute percentage error between the MAP estimates, $\hat{\mu}$ and $\hat{\sigma}$, and the true values. Table 3.1 lists the error averages over all 500 sequences for the different parameter combinations after 0 (i.e., the percentage error between the modes of the chosen priors and the respective true values), 5, 10, 15 and 20 questions. My simulation design allows me to analyze how many questions are necessary to reliably recover parameters, how parameters deteriorate as the response error η increases, and how influential my priors are on the final estimates. Table 3.1 contains answers to all of these questions.

Number of questions asked: As expected, for all parameter combinations, the parameter estimates become more accurate as the number of questions increases. Average absolute percentage error decreases steadily along a sequence of questions. Inspecting the upper half of Table 3.1 (for lower values of η) shows that average deviations from the true values are as low as 4.3% after only 10 questions and go on to decrease to just 2.9% after the 20th question. Even when prior deviations are close to 60%, the average error after question 20 is a mere 4%. Estimates thus converge to the true values rather quickly.

Response error (η): The lower the response error, the higher the information contained in a single binary judgment. Thus, my estimates are expected to be more accurate for simulated respondents with low η values than for those with higher values. This is exactly what Table 3.1 shows. All parameter combinations with lower η (0.05) have, at each point in the sequence of questions, (on average) MAP estimates that are closer to the true values than the same parameter combinations with higher η (0.15). Further support comes from Figure 3.3 illustrating the evolution of MAP estimates as a function of the number of questions asked. The mean MAP estimate at each point along the sequence of questions is enclosed by a 95% confidence interval over the 500 simulation runs. Comparisons of Figure 3.3(a) with 3.3(b) and Figure 3.3(c) with 3.3(d) illustrate that while mean MAP estimates have only marginally different evolution paths for varying η values, the 95% confidence intervals for higher response error rates are obviously wider. Figure 3.4 illustrates that this effect is even stronger for cases with a larger difference between prior modes and true values.



(a) Evolution for $\hat{\mu}$ with $\eta = 0.05$.

(b) Evolution for $\hat{\mu}$ with $\eta = 0.15$.

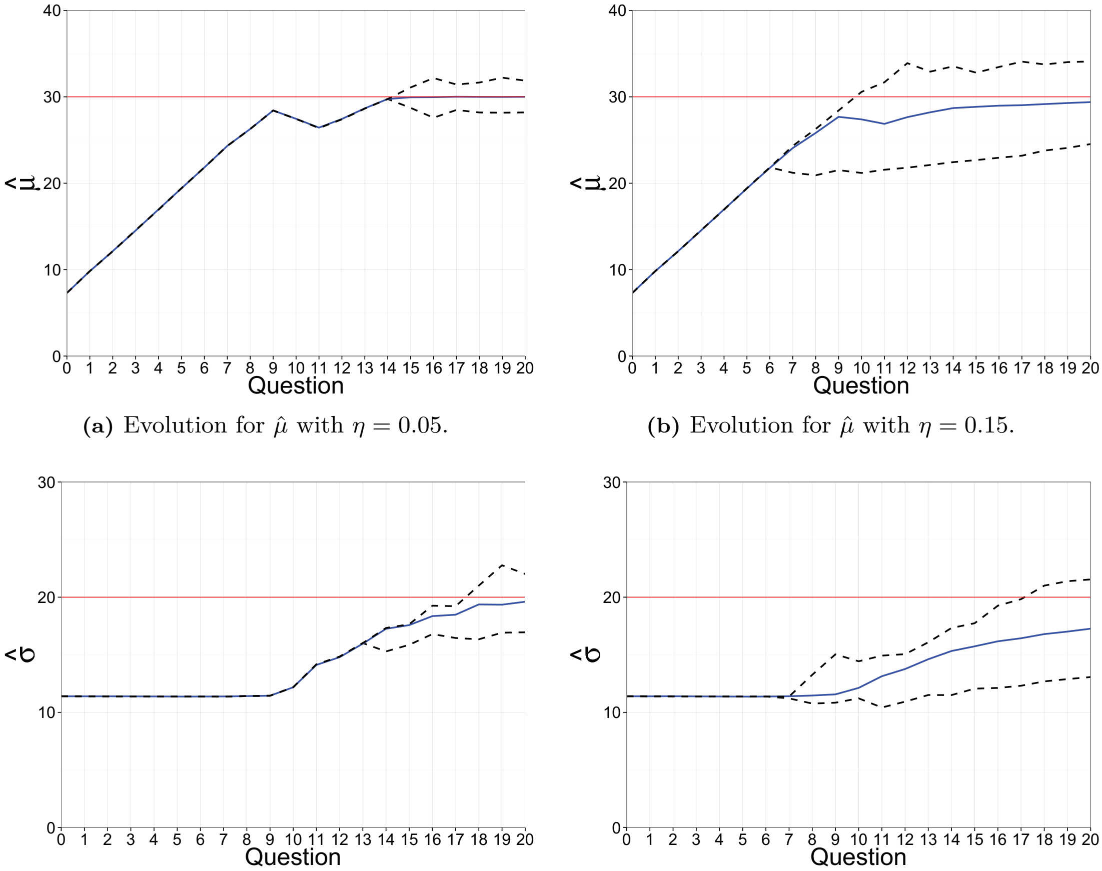
(c) Evolution for $\hat{\sigma}$ with $\eta = 0.05$.

(d) Evolution for $\hat{\sigma}$ with $\eta = 0.15$.

Figure 3.3: Simulation results with $\mu = 10$ and $\sigma = 12$ for different values for η. The thin solid red line corresponds to the true value and the thick solid blue line illustrates the evolution of the average MAP estimate. The dotted black lines define the 95% confidence intervals for the 500 sequences.

Prior sensitivity: For some of my parameter specifications, true values and prior modes are quite close, e.g., with $\mu = 10$ ($\sigma = 12$) and the prior mode at 7.3 (11.4). This is not an unrealistic scenario and it is, therefore, also worth examining how parameter estimates evolve in these cases. However, to test how an increased difference between prior values and true values adds to the parameter evolution, I chose parameter combinations with an absolute average percentage error of the prior ranging from 40% to 60%. Table 3.1 and Figure 3.4 show that, even in such cases, MAP estimates rapidly converge to the true values with maximal average deviations of just 4% after 20 questions. The adaptive procedure first determines the optimal MAP estimate $\hat{\mu}$ and, having accomplished this, goes on to detect an estimate for the true σ.[64] As expected, the further away the true μ is from the prior mode, the longer it takes for my procedure to settle at values of low error as comparisons between Figure 3.3(a) with 3.4(a) and Figure 3.3(c) with 3.4(c) illustrate. However, while estimates are further away from the true values after up to 10 questions compared to cases with low prior-true value difference, 20 questions are enough to reach average absolute percentage error values in the range of just 4%.

[64] This appears to be due to how the chosen norm of the Hessian of the underlying model weights mean and standard deviation parameters.



(a) Evolution for $\hat{\mu}$ with $\eta = 0.05$.

(b) Evolution for $\hat{\mu}$ with $\eta = 0.15$.

(c) Evolution for $\hat{\sigma}$ with $\eta = 0.05$.

(d) Evolution for $\hat{\sigma}$ with $\eta = 0.15$.

Figure 3.4: Simulation results with $\mu = 30$ and $\sigma = 20$ for different values for η. The thin solid red line corresponds to the true value and the thick solid blue line illustrates the evolution of the average MAP estimate. The dotted black lines define the 95% confidence intervals for the 500 sequences.

3.5.5 Random Question Selection

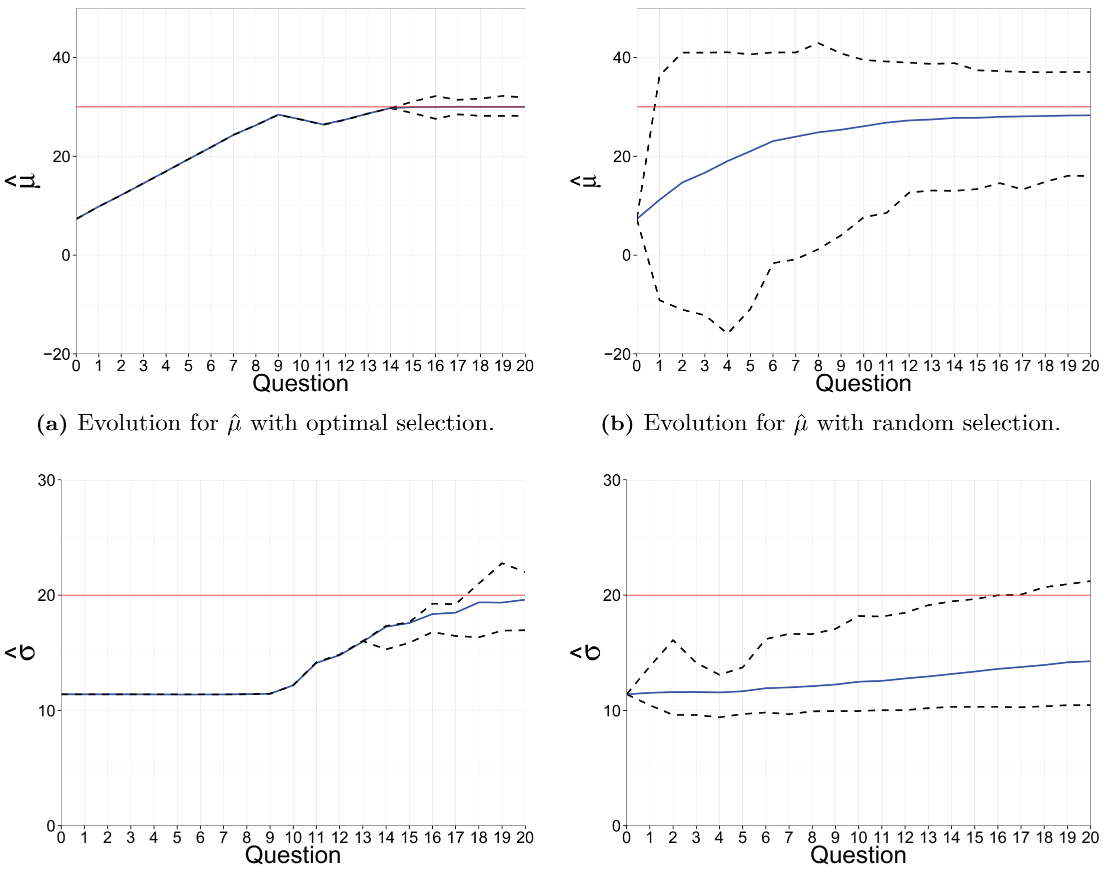
Parameter specification			Number of questions asked				
μ	σ	η	0	5	10	15	20
10	12	0.05	0.160	0.268	0.221	0.181	0.156
10	20	0.05	0.350	0.506	0.452	0.416	0.353
30	12	0.05	0.403	0.210	0.126	0.101	0.087
30	20	0.05	0.593	0.399	0.304	0.254	0.221
10	12	0.15	0.160	0.335	0.297	0.245	0.224
10	20	0.15	0.350	0.567	0.518	0.452	0.386
30	12	0.15	0.403	0.212	0.140	0.125	0.126
30	20	0.15	0.593	0.410	0.343	0.309	0.282

Table 3.2: Average absolute percentage error between simulated MAP estimates and true values for randomly chosen questions.

Furthermore, I investigate how effective my optimal question selection procedure is at increasing information content relative to a simple random selection. Randomly selected here means that

I sample from the same grid of intervals (described in Subsection 3.5.1) at random with equal probability and without replacement.

Table 3.2 reports the error averages over all 500 sequences for the different parameter combinations after 0, 5, 10, 15 and 20 questions if questions are randomly selected. Compared with the values for the optimal question selection procedure listed in Table 3.1, it shows that for all parameter specifications and at each point of the sequence of questions, deviations are much larger for random sampling than for optimal selection with percentage deviations of up to 38% after 20 answered questions.



(a) Evolution for $\hat{\mu}$ with optimal selection.

(b) Evolution for $\hat{\mu}$ with random selection.

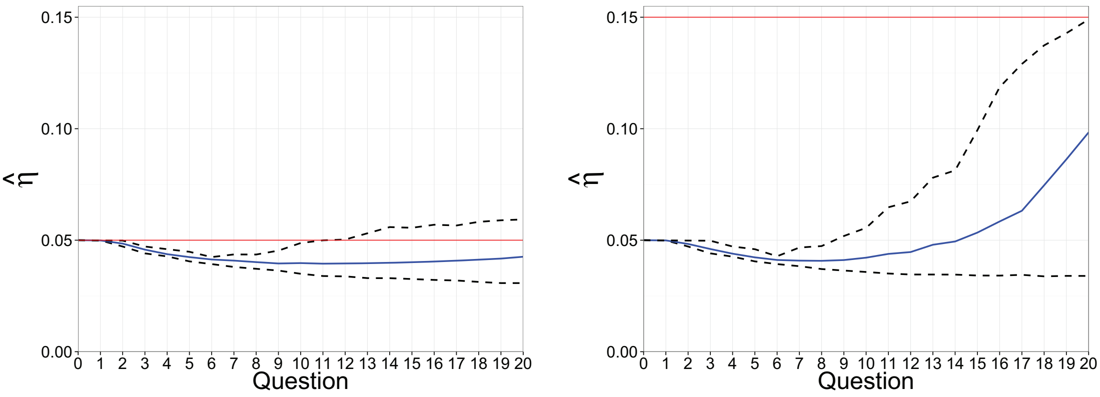
(c) Evolution for $\hat{\sigma}$ with optimal selection.

(d) Evolution for $\hat{\sigma}$ with random selection.

Figure 3.5: Simulation results with $\mu = 30$ and $\sigma = 20$ for different question selection procedures. The thin solid red line corresponds to the true value and the thick solid blue line illustrates the evolution of the average MAP estimate. The dotted black lines define the 95% confidence intervals for the 500 sequences.

In support of this, Figure 3.5 illustrates $\hat{\mu}$ and $\hat{\sigma}$ evolution paths for optimal and random question selection. As shown, when intervals are chosen randomly, MAP parameter estimates respond to answers but move only slowly away from their prior modes. This is because most questions concern intervals that contain almost identical probability under the subjective distribution and so, given the response error, little is learned from the respondent's judgment. Under the adaptively optimal setup, MAP parameter estimates move rapidly away from their prior modes and toward the true parameter values. Moreover, the confidence intervals are considerably narrower when questions are chosen adaptively rather than when they are chosen randomly.

3.5.6 Response Error Recovery



(a) Evolution for $\hat{\eta}$ with true $\eta = 0.05$.

(b) Evolution for $\hat{\eta}$ with true $\eta = 0.15$.

Figure 3.6: Simulation results for different values of η. The thin solid red line corresponds to the true value and the thick solid blue line illustrates the evolution of the average MAP estimate. The dotted black lines define the 95% confidence intervals for the 500 sequences.

In the previous subsections, the focus is solely on the two parameters μ and σ, since these are mainly of interest to researchers working with survey data. While the error parameter η does not give any direct information about the shape of the individual belief distribution, it is of great value for identifying anything from responses that are strictly consistent with the laws of probability to responses that are made entirely at random.

Correctly eliciting the error parameter η is a challenging task as it attempts to estimate the probability of a tail event with only few observations. This problem is evident in Figure 3.6, showing that the recovery MAP estimates $\hat{\eta}$ over the course of the 20 questions in simulations are strongly influenced by the prior with a tendency toward zero after very early questions.[65] This is not surprising since, as only few questions are asked, only few errors can be made.[66] Additionally, questions become more difficult to answer over the course of the 20 questions resulting in potentially more informative choices for the estimation of η at later questions. At the end of the 20 questions, however, for both cases, $\eta = 0.05$ and $\eta = 0.15$, the 95% confidence intervals around the mean MAP estimator become larger and enclose the true values (Figure 3.6).

3.6 The Online Study

I chose Amazon's Mechanical Turk (MTurk) to conduct my online-experiment. MTurk is an online-platform where online workers complete web-based tasks anonymously for small sums of money. While the platform was originally created for simple classification tasks, it is becoming increasingly popular among social scientists for conducting experiments. Its advantages over conventional lab studies are numerous and include access to a large subject pool at low cost, a sample population closer to (though still far from) the general population than students in a lab, an allowance for

[65] The simulation procedure was the same as for the above parameter recovery analysis. The true μ and σ parameters are equal to 30 and 20, respectively, with none of the conclusions qualitatively changing for other settings.

[66] A related issue can be found in Wang et al. (2010) for the recovery of precision parameters in simulations.

a rapid iteration of design, and the possible exploitation of the geographic dispersion of subjects. However, the platform is not without drawbacks. Relative to conventional lab studies, some aspects of the experimental environment are out of the experimenter's control. Moreover, the useable incentive schemes are not as flexible as in the lab, and related to loss of control, some of the participants at MTurk might be trained "lab rats" conditioned to get through the experiment as quickly as possible. However, despite the many differences to controlled lab experiments, many results from classical lab studies from the judgment and decision-making literature have been successfully replicated using MTurk (see, e.g., Paolacci et al. (2010) and Crump et al. (2013)).
Since my elicitation mechanism's (main) target subject group are economic and mathematical laymen answering survey questions, I chose MTurk for my study as its subject pool is closer to my target group than academically trained students in the lab.

3.6.1 The Domains

In my online study I probed respondents' subjective beliefs by asking 20 optimally chosen questions about four different domains:

- **DJIA**: the return in percent on the Dow Jones Industrial Average until years' end,
- **Weather**: the daily high temperature in Fahrenheit in New York City's Central Park at years' end,
- **Gasoline Price**: the price in US$ of one gallon of regular gasoline at local gas stations at years' end,
- **Diesel Price**: the price in US$ of one gallon of regular diesel at local gas stations at years' end.

I chose these domains because they differ systematically along several dimensions. Some domains are closer to people's daily lives, some more remote but nonetheless important (e.g., the weather vs. stock market returns); in some domains informational feedback is frequent while in others it is less frequent (e.g., the prices of gasoline vs. diesel for people who only regularly fill up on one of the two). The kinds of questions I asked about the DJIA domain (and analogously about the other domains) were then of the form:

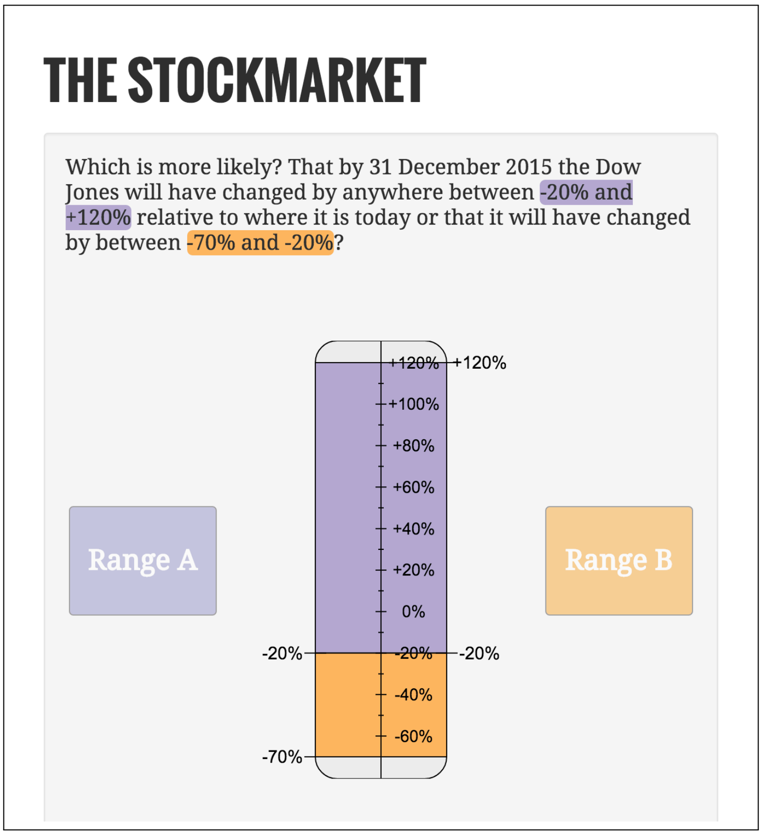
"Which is more likely? That by 31 December 2015 the Dow Jones Industrial Average will have changed by anywhere between -20% and +120% relative to where it is today or that it will have changed by between -70% and -20%?"

Using the above notation from Subsection 3.3.1 this would translate into interval event $A = [A_l, A_u] = [-20, +120]$ and interval event $B = [B_l, B_u] = [-70, -20]$ between which a choice has to be made.
The interval pairs for the DJIA domain with which the subjects are confronted in the experiment are generated from a grid as described above in Subsection 3.5.1. The grid for the Weather domain ranges between [-25°F ; +100°F] with a step size of 5 resulting in 17,550 non-overlapping event pairs from which the optimal binary judgments are picked. For both the Gasoline and the Diesel Price domain the grid ranges between [US$1.5 ; US$4.5] with a step size of 0.1. This results in 35,960 event pairs to be evaluated at each knot of the question-tree.
I used historical data to construct the distributions capturing my prior beliefs. For all four domains the respective μ prior is set to be normal and the other parameters, σ and η, are inverse gamma distributed. These distributions' parameters are specified such that the μ priors are centered at the respective domain specific historical mean and the σ priors have their modal values at the

historical standard deviation. The η priors are the same for all four domains with the mode at 0.05, just like in the simulation study and as described and explained in Subsection 3.5.1.[67]
In addition to the 20 optimally chosen questions, for each domain, I randomly (i.e., at random points in the sequence and in random order) inserted 10 questions in which the two events partition the state space (i.e., their union covers the entire state space). While one event interval increases through the 10 questions covering an ever larger share of the entire state space the other interval decreases accordingly. I use the answers to these partition questions below in order to perform a sanity check on participants' responses by looking at the degree of multiple switching.[68]

3.6.2 Design and Implementation

After accepting my online offer, each MTurk participant was shown a welcome page from which she could move at her own pace (by clicking back and forth buttons) through detailed written and visual instructions of the coming tasks.[69] After confirming that the instructions were understood, participants were then asked to make their judgments about all four domains, the order of the domains having been randomized between respondents.



THE STOCKMARKET

Which is more likely? That by 31 December 2015 the Dow Jones will have changed by anywhere between -20% and +120% relative to where it is today or that it will have changed by between -70% and -20%?

Figure 3.7: Elicitation screen for the DJIA domain.

As illustrated in Figure 3.7, for each binary judgment participants see an elicitation screen with the respective question in written words at the top of the screen and visualized below as a vertical

[67] Figures illustrating the respective prior densities can be found in Appendix A.3.3.
[68] For computational reasons partition questions were not used to compute next optimal questions.
[69] Screen-shots of the instructions are shown in Appendix A.3.5.

number line with the two possible choices emphasized through colors. In order to make their choices, participants had to click with their cursor on one of the two buttons, *Range A* or *Range B*, which was linked to the respective interval on the number line by its color. The link between the upper/lower interval on the number line and the left/right button was randomly permuted after each given answer. In addition, I imposed a 5 second waiting time after each choice to make it impossible for subjects to rush through the questions by randomly clicking on the buttons.
63 participants are recruited via MTurk, each was paid a flat rate of US$4 for the completion of the four questionnaires within 60 minutes after accepting my offer.[70] All participants were US residents (according to their IP-addresses).

3.6.3 Experimental Results

My approach of analyzing the experimental data is threefold. First, I use answers to survey questions and the completion time of the experimental tasks to assess the user-friendliness of the experimental design. Second, I adopt simple statistical testing to sanity check the responses. This is done by utilizing the concept of multiple switching and exploiting specific features of the experimental design. Third, I estimate respondents' subjective belief parameters by means of different econometric approaches, i.e., I employ conventional, recursive, and hierarchical Bayesian procedures. The resulting estimators then help me to check if and how well the elicited beliefs and exogenous related measures are correlated.

3.6.3.1 User-Friendliness

The average time it took respondents to read the instructions and answer all 120 questions (20 optimally chosen and 10 partition questions for each of the four domains) was 25 minutes, with a minimum of 15 and a maximum of 45. That is, eliciting a participant's subjective beliefs about a single domain with 20 optimal questions (without the 10 non-optimal partitions) would have taken roughly 4 minutes on average. In field-experiments and large-scale surveys the time to complete tasks or fill out questionnaires is often a concern since tedious experimental/survey designs might increase non-response rates and reduce the reliability of the given answers due to bored and easily distracted participants (see, e.g., Deutskens et al. (2004) and Galesic and Bosnjak (2009)). The relatively short time it takes to complete my elicitation tasks is a convenient feature of the method that can be of great value in computer based surveys where participants can opt out easily.

[70] My design and choice of domains (with realizations far in the future) preclude me from making payment conditional on the realization of the elicited events as is often done in lab experiments. Some may regard this as a serious disadvantage and worry about the lack of incentives for truth-telling leading to bias. Most evidence in the literature on belief elicitation suggests, however, that this worry is unfounded and that a lack of incentives results not in bias but merely in more noisy responses (see Armantier and Treich (2013) and Trautmann and van de Kuilen (2015)). See also the detailed discussion of how to foil strategizing respondents in an adaptive approach like mine in Wang et al. (2010), with a summary provided in Appendix A.3.1.

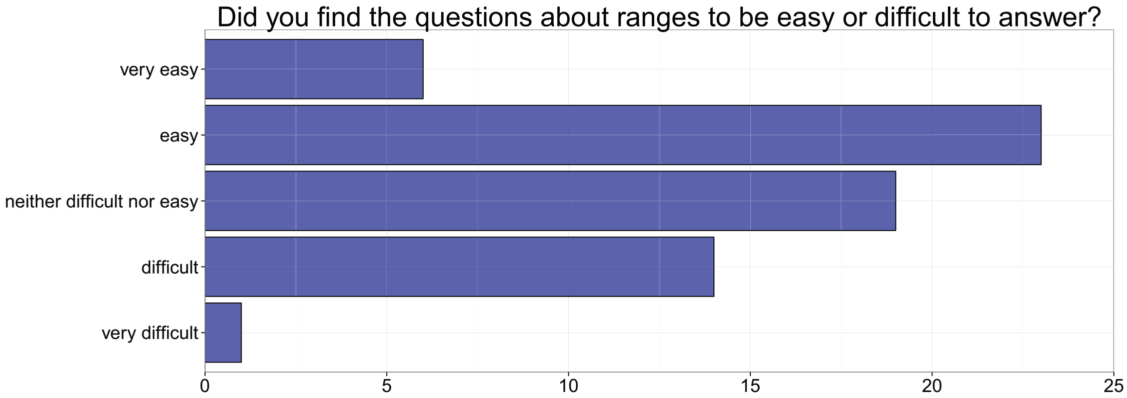
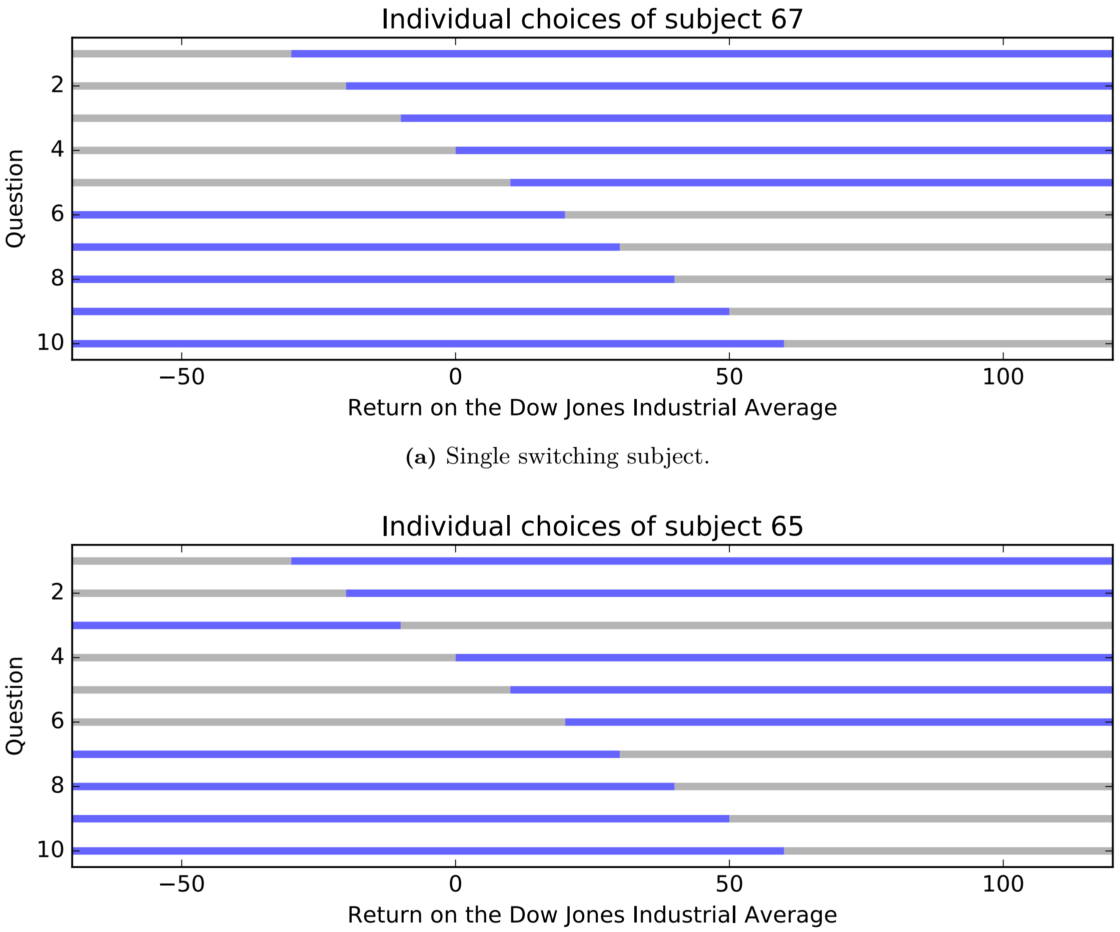


Figure 3.8: Number of experimental subjects giving a categorized answer.

After having completed the elicitation tasks, I asked the respondents about their perceived difficulty of the questions (see Figure 3.8). Only about 20% stated that my design was “difficult” or “very difficult”. Most participants perceived the tasks as easy. Overall, the short time it took the subjects to complete my questionnaire and the low perceived difficulty both indicate a satisfactory level of user-friendliness.

3.6.3.2 Sanity Checks

Multiple switching: An indication of whether or not the participants understood my design and were able to give consistent answers is the degree of multiple switching on the randomly inserted partition questions, a concept I borrow from the literature on multiple price lists designs in risk preference elicitation (see, e.g., Holt and Laury (2002)).
Figure 3.9 illustrates that with each task over the ten partitions the right interval decreases in length while the left interval simultaneously increases in length. According to the laws of probability (see Kolmogorov (1933)), if a participant starts with choosing the (larger) right interval for the first couple of partitions and at a certain partition switches to the ever-growing left interval then she should chose the left interval for all remaining partitions (such as experimental subject 67: Figure 3.9(a)). Hence, participants should not switch back and forth between left and right over the 10 partitions (as experimental subject 65 did: Figure 3.9(b)).
In my online-experiment, over all subjects and all four domains, the multiple switching rate for the partition questions is a surprisingly low 10%. This is a strikingly low number, especially because, unlike in a multiple price list, these questions did not appear in order but were inserted at random points in the sequence and with random ordering so that the problem faced by the participants is not as easy as Figure 3.9 might suggest.



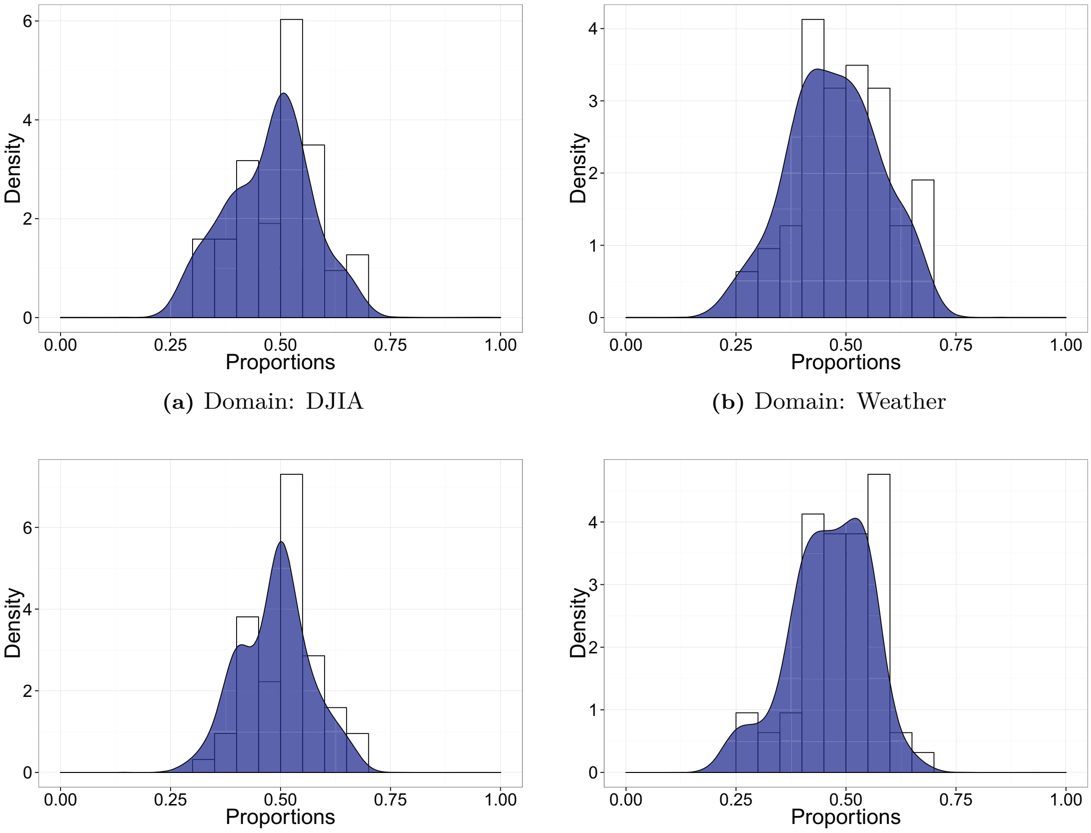
(a) Single switching subject.

(b) Multiple switching subject.

Figure 3.9: Exemplary partition switching behavior in the DJIA domain for two online-participants. Blue colored intervals are the ones chosen.

Heuristic responses: Another indication of whether or not the participants understood the design and actively participated in the experiment is the degree to which they resorted to response patterns that are independent of the specific question asked. Since the link between the upper/lower interval on the number line and the left/right button was randomly permuted after each given answer, if the participants made active judgments they should exhibit proportions of either "left" or "right" choices that are not significantly different from 50%. If, on the other hand, they simply always clicked on, e.g., the left button to get through the experiment as fast as possible this would show up as a proportion of "left" choices significantly higher than 50%.
Figure 3.10 shows the densities for the proportions of "left" choices for my sample of participants, separately for each of the four domains. Throughout all four domains, most of the probability mass is centered in a small neighborhood around proportions of 50% suggesting an active participation in the experiment. This first impression is further supported by the results of binomial tests. For 96% of all participants a binomial test cannot reject the null that the proportion of "left" choices is 50% at a significance level of 5%.[71] As a further test for heuristic responses, unless people consciously pattern responses, choosing "left" should not be auto-correlated. Using a runs-test for binary time series (O'Brien and Dyck, 1985), I cannot reject zero auto-correlation at 5% significance for 92.5% of the sequences in my data.

[71] This failure to reject is no mere consequence of a low number of observations. Pooling the data over all four domains raises the number of judgments from 30 to 120, but leads to similar conclusions.



(a) Domain: DJIA

(b) Domain: Weather

(c) Domain: Gasoline Price

(d) Domain: Diesel Price

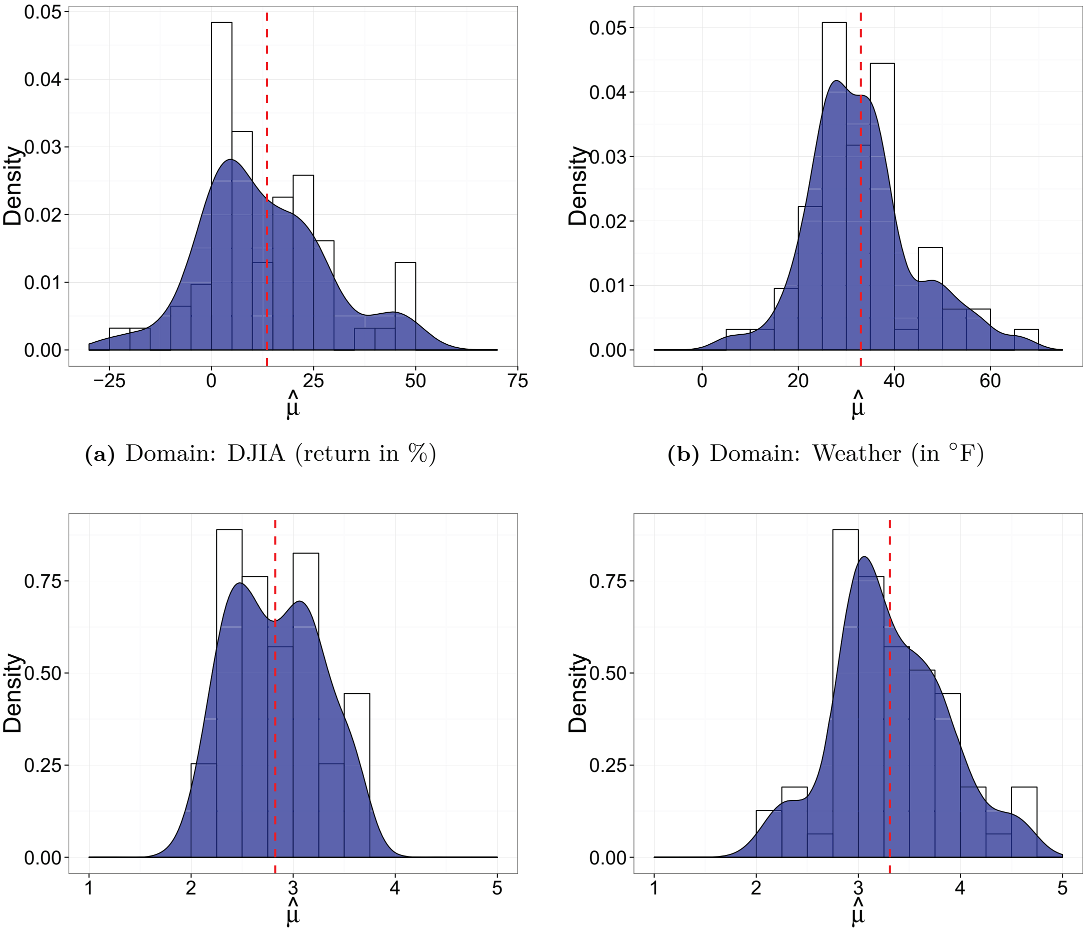
Figure 3.10: Densities for the proportions of "left" choices for all four experimental domains.

3.6.3.3 Estimation Results

The elicited beliefs in terms of MAP parameter estimates (again marked by ˆ) are estimated from the experimental data following the approach described in Subsection 3.3.2.1. I estimate the parameters using a conventional and a recursive technique.[72] Here and in what follows, since only few estimators differed between the two approaches, but the recursive technique supplied higher likelihood values, its estimators are reported (none of the conclusions would change qualitatively using the conventional estimates).

As demonstrated in the previous subsection, experimental participants behaved in large parts consistently (i.e., few multiple switching) and made not too extreme choices (i.e., few heuristic responses). This is also reflected in Figure 3.11 illustrating domain specific $\hat{\mu}$-densities for the whole subject pool.

[72] A recursive technique sequentially estimates the posterior distribution by using the incoming observations over time to update the MAP estimates. The conventional approach simply pools all observations to construct the likelihood.



(a) Domain: DJIA (return in %)

(b) Domain: Weather (in °F)

(c) Domain: Gasoline Price (in US$/G)

(d) Domain: Diesel Price (in US$/G)

Figure 3.11: Subject pool densities of MAP estimates $\hat{\mu}$, with dashed lines illustrating the domain specific mean.

DJIA: 90% of the probability mass of the $\hat{\mu}$-density for the expected percentage changes of the DJIA until year's end range between -10% and +45% with a mean of 13.5%.[73] The distribution is only marginally skewed to the right with the median (9.9%) being just slightly smaller than the mean.

Weather: For the daily high temperature in New York City's Central Park at years' end, the $\hat{\mu}$-density is almost symmetric with mean (33°F) and median (32°F) close together. Only 10% of the $\hat{\mu}$ in the weather domain fall outside the interval [19°F ; 53°F].[74]

Gasoline Price: The $\hat{\mu}$-density for beliefs about the gasoline price at years' end is quite symmetric with both mean and median at 2.8 US$/G. 90% of the elicited $\hat{\mu}$ parameters cover a range between 2.2 US$/G and 3.5 US$/G.[75]

Diesel Price: Similar to the Gasoline domain, the $\hat{\mu}$-density for the beliefs about the diesel price is fairly symmetric with only a small difference between mean (3.3 US$/G) and median (3.2 US$/G).

[73] Just as a reference, the actual yearly DJIA returns of 2015(-2.2%), 2014(7.5%), 2013(26.5%), 2012(7.2%), and 2011(5.5%) are well within that interval (source: own calculations). Note, however, that the experiment was conducted in early 2015 so that the data reflect beliefs about the DJIA's 9 months performance rather than its annual returns.

[74] The actual daily high temperatures in New York City's Central Park at years' end of the last five years were 45°F (2016), 48°F (2015), 32°F (2014), 32°F (2013), and 37°F (2012) (source: National Weather Service, `http://www.weather.gov/`).

[75] Gasoline prices vary largely between US states. The US wide average fluctuated between 1.69 US$/G and 3.93 US$/G over the last 5 years. On 31 December 2015 this average was at 2.00 US$/G and on the day of the experiment it was at 2.43 US$/G (source: `http://then.gasbuddy.com`).

The 90% range of the density covers values from 2.3 US$/G up to 4.2 US$/G. The larger range (covering higher prices) for diesel price beliefs compared to gasoline price beliefs reflects the fact that diesel is usually more expensive than regular gasoline in the US.[76]

Analog domain specific densities of the MAP estimates $\hat{\sigma}$ and $\hat{\eta}$ also show reasonable values. Median $\hat{\sigma}$ parameter values are 11.0% (DJIA), 7.5°F (Weather), 0.37 US$/G (Gasoline Price), and 0.38 US$/G (Diesel Price). Median values of the $\hat{\eta}$-densities are 0.058 (DJIA), 0.045 (Weather), 0.047 (Gasoline Price), and 0.053 (Diesel Price). The corresponding distributions are illustrated in Appendix A.3.4.

3.6.3.4 Plausibility Checks

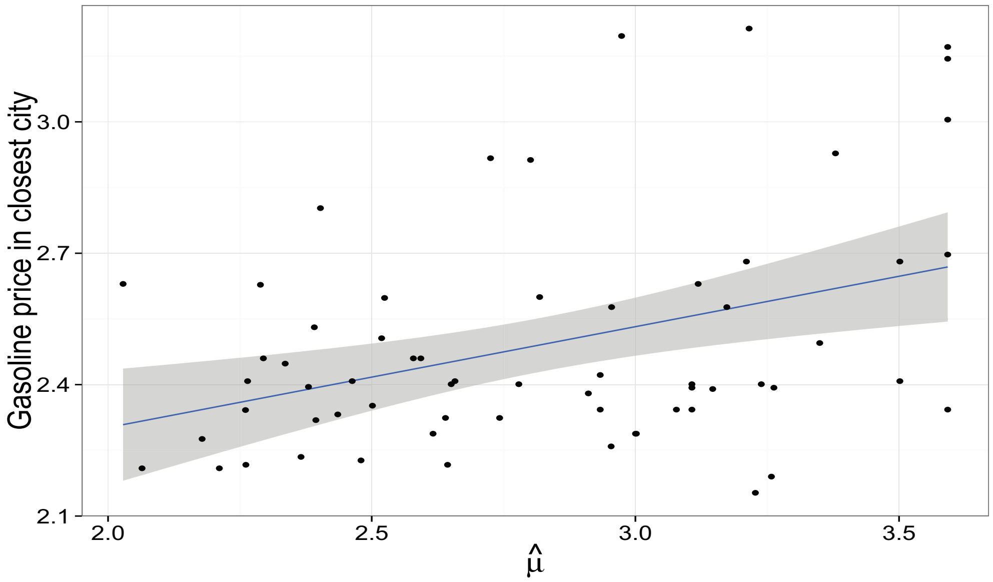


Figure 3.12: Correlation between gasoline price beliefs (in terms of MAP estimates $\hat{\mu}$) and current prices in US$ (regression parameter value: 0.644, p-value < 0.05).

I also check for correlations of the elicited beliefs with other related measures.

Gasoline Price: One relation that seems especially appropriate for this purpose is the one between the respondents' subjective beliefs about the future gasoline price and the actual price in the closest city at the time of the experiment.[77] A positive correlation here would deliver further support for the adaptive design.

For my analysis, I obtain the latter measure by using the geo-coordinates derived from respondents' IP addresses as place of residence proxy. For each place of residence I then picked the closest city and scraped the corresponding gas price listed for the area around that city for the time of the experiment from the largest US web-database concerning gas prices.[78] Figure 3.12 illustrates the respective relation. It shows a positive correlation at a significance level of <5% (regression parameter value: 0.644). That is, respondents living near a city with higher gasoline prices have (on average) higher subjective beliefs about this price.

[76] On 31 December 2015 the US wide average diesel price was at 2.26 US$/G and on the day of the experiment it was at 2.78 US$/G (source: `http://then.gasbuddy.com`).

[77] While planning the experiment I expected the number of those driving gasoline and diesel cars to be quite balanced in the US population, which is not the case in reality. This is also reflected in my experimental sample, where only 1 of 63 subjects regularly fills up on diesel fuel. Therefore, I focus on beliefs about gasoline prices in the following.

[78] Supplying historical and real-time data on gas prices, `http://then.gasbuddy.com` was my data source.

DJIA: Another relation that is well suited to test for a reasonable correlation is the one between the respondents' subjective beliefs about the future DJIA performance and whether or not they were actively trading and have put more money in the stock market in the weeks just before the experiment. Those who trade actively and put more money in the market (roughly 10% of my experimental sample) act as if they were optimistic about the future performance of the market and this should also be reflected in their beliefs about the future DJIA evolution, i.e. they should expect higher percentage increases of the DJIA.[79]

This is exactly what I find using the estimated MAP parameters ($\hat{\mu}$) and data collected from the experimental survey in a logit regression. Those who act optimistically exhibit beliefs about future DJIA percentage changes that are, on average, more positive than the beliefs of those who did not increase their investment in the weeks just before the experiment (parameter value: 0.039, p-value < 0.1).

Weather: In order to test for a reasonable correlation in the weather domain I again use the geo-coordinates of my subjects' IP-addresses. That is, I check if the distance between a subject's place of residence and New York City's Central Park (in miles), is positively related to the individual response error MAP parameter, $\hat{\eta}$. I expect those living further away from the Central Park to exhibit larger $\hat{\eta}$ parameter estimates in the weather domain. Due to small sample issues, I employ the hierarchical estimation approach described in Subsection 3.3.2.2.

The results show a significant positive correlation between a subject's distance to New York City's Central Park and the individual MAP parameter $\hat{\eta}$ (regression parameter value: 6.41e-06, p-value < 0.1).

3.7 Conclusion

The success of examining the links between beliefs and decision making, may it be in areas such as economics (e.g., insurance, investment or savings decisions), health (e.g., cigarette consumption or vaccination) or any other (e.g., drunk driving), depends crucially on mechanisms that allow an accurate elicitation of such beliefs in a user-friendly and timely manner.

I present a new elicitation method that makes use of an adaptive approach maximizing statistical efficacy and minimizing effort for the respondents. By adaptively choosing questions, such that their answers contain maximal information, in a well-defined statistical sense, the estimates are less likely to be distorted by the respondents' inherent or contextual disturbances. Furthermore, the experimental design minimizes the time and cognitive resources expended by the respondents. That is, they only have to answer a limited number of very simple binary choice questions, which should not only mitigate resource depletion effects (see, e.g., Schmeichel et al. (2003)) but also helps to overcome the usual methodical problems of belief elicitation, e.g., inconsistent answers that violate the laws of probability or bunching at certain probability statements (see, Delavande and Rohwedder (2008) and Binswanger and Salm (2013)).

Other advantages of the presented method are a possible pre-screening of respondents since Bayesian posterior modes are calculated after each given answer and the variability with which my methodology can be applied to different domains and probabilistic structures of almost arbitrary complexity. In combination, these features of my method make studies with a large online and/or laymen population (in contrast to the commonly used student subject pools in economic experiments) possible and more practical.

The results of my simulation study suggest that my binary approach is well suited to recover the true parameters and does so with surprisingly few observations. As expected, I find that (i)

[79] I did not ask my subjects about their portfolio just whether or not they bought more stocks. Thus, it could well be that some of those who put more money in the market recently shorted against the DJIA, which would question my above reasoning. However, based on the professional and educational background of my subject pool I assume this is not the case.

my estimates are more accurate if I increase the number of questions; (ii) my estimates are less accurate if I increase the noisiness of the simulated respondents; and (iii) the difference between prior mean/mode and true values only marginally influences estimation accuracy after 20 questions. I also conducted an online-experiment, the results of which support the user-friendliness of the adaptive design with comparatively short completion times, low self-reported difficulty, and very low degrees of multiple switching and heuristic responses. Additionally, the elicited belief parameters show reasonable values and I could show reasonable and significant correlations between the elicited beliefs and related measures implying a satisfactory goodness of elicitation.

Bibliography

K. Abbink and B. Rockenbach. Option pricing by students and professional traders: A behavioral investigation. *Managerial and Decision Economics*, 27:497–510, 2006.

M. Abdellaoui, A. Baillon, L. Placido, and P. Wakker. The rich domain of uncertainty: Source functions and their experimental implementation. *American Economic Review*, 101:695–723, 2011.

J. Abernethy, T. Evgeniou, O. Toubia, and J.P. Vert. Eliciting consumer preferences using robust adaptive choice questionnaires. *IEEE Transactions on Knowledge and Data Engineering*, 20:145–155, 2008.

A.M. Ahmed, L. Andersson, and M. Hammarstedt. Are gay men and lesbians discriminated against in the hiring process? *Southern Economic Journal*, 79:565–585, 2013.

D.J. Aigner and G.G. Cain. Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review*, 30:175–187, 1977.

G.M. Allenby and P.E. Rossi. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89:57–78, 1998.

S. Andersen, J. Fountain, G.W. Harrison, and E.E. Rutström. Estimating subjective probabilities. *Journal of Risk and Uncertainty*, 48:207–229, 2014.

O. Armantier and N. Treich. Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging. *European Economic Review*, 62:17–40, 2013.

L. Arrondel, H. Calvo-Pardo, and D. Tas. Subjective return expectations, information and stock market participation: Evidence from france. *Paris School of Economics- Banque de France and University of Southampton Working Paper*, pages 1–59, 2014.

K. Arrow. The theory of discrimination. *Discrimination in Labor Markets*, 3:3–33, 1973.

R.B. Avery, P.S. Calem, and G.B. Canner. Consumer credit scoring: Do situational circumstances matter? *Journal of Banking and Finance*, 28:835–856, 2004.

G. Azmat and R. Ferrer. Gender gaps in performance: Evidence from young lawyers. *Journal of Political Economy (Forthcoming)*, 2015.

M. Bagues and M.J. Perez-Villadoniga. Why do i like people like me? *Journal of Economic Theory*, 148:1292–1299, 2013.

A. Baillon. Eliciting subjective probabilities through exchangeable events: an advantage and a limitation. *Decision Analysis*, 5:76–87, 2008.

N. Barasinska. Does gender affect investors' appetite for risk? evidence from peer-to-peer lending. *DIW Berlin Discussion Paper No. 1125*, 2011.

N. Barasinska and D. Schäfer. Is crowdfunding different? evidence on the relation between gender and funding success from a german peer-to-peer lending platform. *German Economic Review*, 15:436–452, 2014.

G.S. Becker. *The Economics of Discrimination.* University of Chicago Press, 1957.

S. Benartzi and R.H. Thaler. Risk aversion or myopia? choices in repeated gambles and retirement investments. *Management Science*, 45:364–381, 1999.

J.M. Bernardo and A.F.M. Smith. *Bayesian Theory.* New York: Wiley, 1994.

S. Berry, R.J. Carroll, and D. Ruppert. Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97:160–169, 2000.

M. Bertrand and S. Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94:991–1013, 2004.

J. Binswanger and M. Salm. Does everyone use probabilities? intuitive and rational decisions about stockholding. *IZA Discussion Paper 7265*, 2013.

T. Brünner, R. Levinsky, and J. Qiu. Preference for skewness: Evidence from a binary choice experiment. *European Journal of Finance*, 17:525–538, 2011.

A.C. Cameron and P.K. Trivedi. *Microeconometrics: Methods and Applications.* Cambridge University Press, 2005.

M. Carlsson and S. Eriksson. Discrimination in the rental market for apartments. *Journal of Housing Economics*, 23:41–54, 2014.

S. Cerroni and W.D. Shaw. Does climate change information affect stated risks of pine beetle impacts on forests? an application of the exchangeability method. *Forest Policy and Economics*, 22:72–84, 2012.

S. Cerroni, S. Notaro, and W.D. Shaw. Eliciting and estimating valid subjective probabilities: An experimental investigation of the exchangeability method. *Journal of Economic Behavior & Organization*, 84:201–215, 2012.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10:273–304, 1995.

G. Charness and M. Rabin. Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117:817–869, 2002.

H. Chen and A.R. Rao. When two plus two is not equal to four: Errors in processing multiple percentage changes. *Journal of Consumer Research*, 34:327–340, 2007.

F. Christandl and D. Fetchenhauer. How laypeople and experts misperceive the effect of economic growth. *Journal of Economic Psychology*, 30:381–392, 2009.

J.R. Cook and L.A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89:1314–1328, 1994.

B. Cornell and I. Welch. Culture, information, and screening discrimination. *Journal of Political Economy*, 104: 542–571, 1996.

M. Costa-Gomes and G. Weizsäcker. Stated beliefs and play in normal-form games. *Review of Economic Studies*, 75: 729–762, 2008.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34: 187–220, 1972.

J.C. Cox, S.A. Ross, and M. Rubinstein. Option pricing: A simplified approach. *Journal of Financial Economics*, 7: 229–263, 1979.

M.J.C. Crump, J.V. McDonnell, and T.M. Gureckis. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE*, 8:e57410. doi:10.1371/journal.pone.0057410, 2013.

B. de Finetti. La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, 7:1–68, 1937.

M. De Groot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

C. Deck and H. Schlesinger. Exploring higher order risk effects. *Review of Economic Studies*, 77:1403–1420, 2010.

A. Delavande and S. Rohwedder. Eliciting subjective probabilities in internet surveys. *Public Opinion Quarterly*, 72: 866–891, 2008.

A. Delavande, X. Giné, and D. McKenzie. Measuring subjective expectations in developing countries: A critical review and new evidence. *Journal of Development Economics*, 94:151–163, 2011.

E. Deutskens, K. De Ruyter, M. Wetzels, and O. Paul. Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters*, 15:21–36, 2004.

J. Dominitz and C. Manski. Perceptions of economic insecurity: Evidence from the survey of economic expectations. *Public Opinion Quarterly*, 61:261–287, 1997.

T. Drerup, B. Enke, and H.-M. von Gaudecker. Measurement error in subjective expectations and the empirical content of economic models. *IZA Discussion Paper 8535*, 2014.

S. Ebert and D. Wiesen. Testing for prudence and skewness-seeking. *Management Science*, 57:1334–1349, 2011.

C.C. Eckel and P.J. Grossman. Loving the long shot: Risk taking with skewed lotteries. *Working paper, Texas A&M University*, 2014.

E.M. Eisenstein and S.J. Hoch. Intuitive compounding: Framing, temporal perspective, and expertise. *Working paper, Johnson Graduate School of Management, Cornell University*, 2005.

L. Ensthaler, O. Nottmeyer, and G. Weizsäcker. Hidden skewness: on the difficulty of multiplicative compounding under random shocks. *Working paper, Humboldt University Berlin*, 2010.

L. Ensthaler, O. Nottmeyer, G. Weizsäcker, and C. Zankiewicz. Hidden skewness: on the difficulty of multiplicative compounding under random shocks. *Working paper, Humboldt University Berlin*, 2013.

Financial Industry Regulatory Authority. *FINRA Regulatory Note 09-31*. Financial Industry Regulatory Authority. Accessed January 11, 2017, http://www.finra.org/sites/default/files/NoticeDocument/p118952.pdf, 2009.

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

K.Y. Fung and D. Krewski. On measurement error adjustment methods in poisson regression. *Environmetrics*, 10: 213–224, 1999.

X. Gabaix, D. Laibson, G. Moloche, and S. Weinberg. Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review*, 96:1043–1068, 2006.

M. Galesic and M. Bosnjak. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73:349–360, 2009.

Y. Gallen. The gender productivity gap. *Northwestern University Working Paper*, 2016.

P.H. Garthwaite, J.B. Kadane, and A. O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100:680–700, 2005.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis.* Chapman and Hall/CRC, 2005.

G. Gigerenzer and R. Selten. *Bounded Rationality: the adaptive toolbox.* Cambridge: Massachusetts: MIT Press., 2001.

U. Gneezy. Probability judgments in multi-stage problems: Experimental evidence of systematic biases. *Acta Psychologica*, 93:59–68, 1996.

U. Gneezy and J. Potters. An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics*, 112:631–645, 1997.

T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

D. G. Goldstein and D. Rothschild. Lay understanding of probability distributions. *Judgment and Decision Making*, 9:1–14, 2014.

F. Gouret and G. Hollard. When kahneman meets manski: using dual systems of reasoning to interpret subjective expectations of equity returns. *Journal of Applied Econometrics*, 26:371–392, 2011.

L. Guiso, T. Jappelli, and D. Terlizzese. Earnings uncertainty and precautionary saving. *Journal of Monetary Economics*, 30:307–337, 1992.

J.W. Hardin, H. Schmiediche, and R.J. Carroll. The regression calibration method for fitting generalized linear models with additive measurement error. *Stata Journal 3.4*, pages 1–11, 2003.

J.K. Haukka. Correction for covariate measurement error in generalized linear model - a bootstrap approach. *Biometrics*, 51:1127–1132, 1995.

J.K. Hellerstein, D. Neumark, and Troske K.R. Wages, productivity, and worker characteristics: Evidence from plant-level production. *Journal of Labor Economics*, 17:409–446, 1999.

M. Herzenstein, U.M. Dholakia, and R.L. Andrews. Strategic herding behavior in peer-to-peer loan auctions. *Journal of Interactive Marketing*, 25:27–36, 2011.

J. Hey and C. Orme. Investigating generalizations of expected utility theory using experimental data. *Econometrica*, 62:1291–1326, 1994.

C.A. Holt and S.K. Laury. Risk aversion and incentive effects. *American Economic Review*, 92:1644–1655, 2002.

S. Huck, T. Schmidt, and G. Weizsäcker. The standard portfolio choice problem in germany. *SOEP Discussion Paper 650*, 2014.

M. D. Hurd and S. Rohwedder. Stock price expectations and stock trading. *NBER Working Paper(17973)*, 2012.

M. D. Hurd, M. van Rooij, and J. Winter. Stock market expectations of dutch households. *Journal of Applied Econometrics*, 26:416–436, 2011.

R. Iyer, E.I. Khwaja, E.F.P. Luttmer, and K. Shue. Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending? *HKS Faculty Research Working Paper Series, John F. Kennedy School of Government, Harvard University*, 2009.

P.M. Jakus, W.D. Shaw, T.N. Nguyen, and M. Walker. Risk perceptions of arsenic in tap water and bottled water consumption. *Water Resource Research*, 2009.

D. Jenkinson. The elicitation of probabilities: A review of the statistical literature. *BEEP Working Paper, University of Sheffield*, 2005.

W. Jiang and B. Turnbull. The indirect method: Inference based on intermediate statistics – a synthesis and examples. *Statistical Science*, 19:239–263, 2004.

L. Kaas and C. Manger. Ethnic discrimination in germany's labour market: A field experiment. *German Economic Review*, 13:1–20, 2012.

S. Kemp. Perception of changes in the cost of living. *Journal of Economic Psychology*, 5:313–323, 1984.

G. Kezdi and R.J. Willis. Stock market expectations and portfolio choice of american households. *University of Michigan, Working paper*, 2008.

M. Kilka and M. Weber. What determines the shape of the probability weighting function under uncertainty. *Management Science*, 47:1712–1726, 2001.

E.U. Klos, A. Weber and M. Weber. Investment decisions and time horizon: Risk perception and risk behavior in repeated gambles. *Management Science*, 51:1777–1790, 2005.

A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung (engl. "Foundations of the Theory of Probability", translated in 1950).* Springer, Berlin, 1933.

D. Kraus. Does borrowers' impatience disclose their hidden information about default risk? *Working Paper, University of Rostock*, 2013.

T. Kricheli-Katz and T. Regev. How many cents on the dollar? women and men in product markets. *Science Advances*, 2:1–8, 2016.

P. Kuhn and K. Shen. Gender discrimination in job ads: Evidence from china. *The Quarterly Journal of Economics*, 128:287–336, 2013.

K.B. Laskey and G. W. Fischer. Estimating utility functions in the presence of response error. *Management Science*, 33:965–980, 1987.

C.N. Lawrence. Accounting for the "known unknowns": Incorporating uncertainty in second-stage estimation. *Texas A&M International University Working Paper*, 2009.

M. Levy and J. Tasoff. Exponential-growth bias and lifecycle consumption. *Journal of the European Economic Association*, page doi: 10.1111/jeea.12149, 2015.

G. Loewenstein and J.S. Lerner. *The Role of Affect in Decision Making, in Handbook of Affective Sciences.* Oxford and New York: Oxford University Press: 619-642, 2003.

R.D. Luce. A probabilistic theory of utility. *Econometrica*, 26:193–224, 1958.

A. Lusardi and O.S. Mitchell. Financial literacy and planning: Implications for retirement wellbeing. *NBER Working Paper 17078*, 2011.

D. McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics, Academic Press, New York*, pages 105–142, 1974.

C.R.M. McKenzie and M.J. Liersch. Misunderstanding savings growth: Implications for retirement savings behavior. *Journal of Marketing Research*, 48:1–13, 2011.

L. Menapace, G. Colson, and R. Raffaelli. Climate change beliefs and perceptions of agricultural risks: An application of the exchangeability method. *Global Environmental Change*, 35:70–81, 2015.

M.W. Merkhofer. A process for technology assessment based on decision analysis. *Technological Forecasting and Social Change*, 22:237–265, 1982.

E. Mills, A.R. Jadad, C. Ross, and K. Wilson. Systematic review of qualitative studies exploring parental beliefs and attitudes toward childhood vaccination identifies common barriers to vaccination. *Journal of Clinical Epidemiology*, 58:1081–1088, 2005.

W.K. Newey and D. McFadden. *Large sample estimation and hypothesis testing.* Handbook of Econometrics, Elsevier Science, Amsterdam, 1994.

J. M. Nunley, A. Pugh, N. Romero, and R.A. Seals. An examination of racial discrimination in the labor market for recent college graduates: Estimates from the field. *Auburn University Working Paper*, 2014.

P.C. O'Brien and P.J. Dyck. A runs test based on run lengths. *Biometrics*, 41:237–244, 1985.

T. Offerman, J. Sonnemans, G. Van de Kuilen, and P.P. Wakker. A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *Review of Economic Studies*, 76:1461–1489, 2009.

D. Pager. The mark of a criminal record. *American Journal of Sociology*, 108:937–975, 2003.

G. Paolacci, J. Chandler, and P.G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment Decision Making*, 5:411–419, 2010.

E.S. Phelps. The statistical theory of racism and sexism. *American Economic Review*, 62:659–661, 1972.

D.G. Pope and J.E. Snydor. What's in a picture? evidence of discrimination from prosper.com. *Journal of Human Resources*, 46:53–92, 2011.

S. Pötzsch and R. Böhme. *The Role of Soft Information in Trust Building: Evidence from Online Social Lending.* Volume 6101 of the series Lecture Notes in Computer Science, 2010.

J. Price and J. Wolfers. Racial discrimination among nba referees. *The Quarterly Journal of Economics*, 125: 1859–1887, 2010.

H. Raiffa. *Decision Analysis.* London: Addison-Wesley, 1968.

F.P. Ramsey. *"Truth and probability" in "The Foundations of Mathematics and Other Logical Essays".* London, 1931.

E. Ravina. Love & loans: The effect of beauty and personal characteristics in credit markets. *Working Paper, Columbia University*, 2012.

D.A. Redelmeier and A. Tversky. On the framing of multiple prospects. *Psychological Science*, 3:191–193, 1992.

P.A. Riach and J. Rich. An experimental investigation of age discrimination in the english labor market. *Annals of Economics and Statistics*, 99:169–185, 2010.

M. Riddel and W.D. Shaw. A theoretically-consistent empirical model of nonexpected utility: An application to nuclear-waste transport. *Journal of Risk and Uncertainty*, 32:131–150, 2006.

D.O. Rooth. Obesity, attractiveness, and differential treatment in hiring: A field experiment. *Journal of Human Resources*, 44:710–735, 2009.

P.E. Rossi and G.M. Allenby. Bayesian statistics and marketing. *Marketing Science*, 22:304–328, 2003.

P.E. Rossi, G.M. Allenby, and R. McCulloch. *Bayesian Statistics and Marketing.* John Wiley& Sons, Ltd: Wiley Series in Probability and Statistics, 2006.

D. Rothschild. Expectations: Point-estimates, probability distributions, confidence, and forecasts. *Microsoft Research & Applied Statistics Center at Columbia Working Paper*, 2011.

P. A. Samuelson. Risk and uncertainty: A fallacy of large numbers. *Scientia*, 98:108–113, 1963.

L.J. Savage. *The Foundations of Statistics.* Wiley, New York., 1954.

Sawtooth Software. *ACA system for adaptive conjoint analysis.* Technical paper, Sawtooth Software, Sequim,WA. Accessed January 11, 2017, http://www.sawtoothsoftware.com/download/techpap/acatech.pdf., 1996.

K.H. Schlag, J. Tremewan, and J.J. van der Weele. A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, 18:457–490, 2015.

B.J. Schmeichel, K.D. Vohs, and R.F. Baumeister. Intellectual performance and ego depletion: Role of the self in logical reasoning and other information processing. *Journal of Personality and Social Psychology*, 85:33–46, 2003.

B. Schneider, S. Parker, and D. Stein. The measurement of loudness using direct comparisons of sensory intervals. *Journal of Mathematical Psychology*, 11:259–273, 1974.

J. H. Sepanski, R. Knickerbocker, and R. J. Carroll. A semiparametric correction for attenuation. *Journal of the American Statistical Association*, 89:1366–1373, 1994.

S. Shiloh, M. Vinter, and M. Barak. Correlates of health screening utilization: The roles of health beliefs and self-regulation motivation. *Psychology & Health*, 12:301–317, 1997.

C.S. Spetzler and C.-A. S. Staël Von Holstein. Probability encoding in decision analysis. *Management Science*, 22: 340–358, 1975.

V. Stango and J. Zinman. Exponential growth bias and household finance. *Journal of Finance*, 64:2807–2849, 2009.

L.A. Stefanski and J.R. Cook. Simulation-extrapolation: The measurement error jackknife. *Journal of the American Statistical Association*, 90:1247–1256, 1995.

M. Stutzer and S. Grant. Misperceptions of long-term investment performance: Insights from an experiment. *Journal of Behavioral Finance and Economics*, 3:1–20, 2013.

R. Thaler. Some empirical evidence on dynamic inconsistency. *Economics Letters*, 8:201–207, 1981.

W. Theseira. Competition to default? racial discrimination in the market for online peer-to-peer lending. *Working Paper, Wharton School of the University of Pennsylvania*, 2008.

O. Toubia, D. Simester, J.R. Hauser, and E. Dahan. Fast polyhedral conjoint estimation. *Marketing Science*, 22: 274–303, 2003.

O. Toubia, E. Johnson, T. Evgeniou, and P. Delquié. Dynamic experiments for estimating preferences: An adaptive method of eliciting time and risk parameters. *Management Science*, 59:613–640, 2013.

S. T. Trautmann and G. van de Kuilen. Belief elicitation: A horse race among truth serums. *Economic Journal*, 125: 2116–2135, 2015.

P. van Santen, R. Alessie, and A. Kalwij. Probabilistic survey questions and incorrect answers: Retirement income replacement rates. *Journal of Economic Behavior and Organization*, 82:267–280, 2012.

W.A. Wagenaar and S.D. Sagaria. Misperception of exponential growth. *Perception & Psychophysics*, 18:416–422, 1975.

W.A. Wagenaar and H. Timmers. Extrapolation of exponential time series is not enhanced by having more data points. *Perception & Psychophysics*, 24:182–184, 1978.

W.A. Wagenaar and H. Timmers. The pond-and-duckweed problem: Three experiments on the misperception of exponential growth. *Acta Psychologica*, 43:239–251, 1979.

S. Wang, M. Filiba, and C. Camerer. Dynamically optimized sequential experimentation (dose) for estimating economic preference parameters. *Working paper, California Institute of Technology, Pasadena*, 2010.

N.D. Weinstein and M.A. Diefenbach. Percentage and verbal category measures of risk likelihood. *Health Education Research*, 12:139–141, 1997.

G. Weizsäcker. Do we follow others when we should? a simple test of rational expectations. *American Economic Review*, 100:2340–2360, 2010.

M. B. Welsh, M. D. Lee, and S. H. Begg. More-or-less elicitation (mole): Testing a heuristic elicitation method. *In Sloutsky, V., Love, B. and McRae, K. (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 493–498, 2008.

A. Winman, P. Hansson, and P. Juslin. Subjective probability intervals: how to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30:1167–1175, 2004.

J.M. Wooldridge. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press, 2002.

B.R.E. Wright, M. Wallace, J. Bailey, and A. Hyde. Religious affiliation and hiring discrimination in new england: A field experiment. *Research in Social Stratification and Mobility*, 34:111–126, 2013.

A. Zussman. Ethnic discrimination: Lessons from the israeli online market for used cars. *The Economic Journal*, 123:433–468, 2013.

A

Appendix

Contents:

A.1 Hidden Skewness

A.1.1 Stylized Experiment in the Introduction

A.1.1.1 Experimental Design and Exponential Growth Bias Prediction

The experimental asset of interest, Security A, follows a binomial-tree -60%/+70% process over 12 periods that is identical to Security A in Study 1(b), with an initial price of 10,000. In three experimental sessions, 69 students at Technical University Berlin are presented with the growth process of Security A and are asked to pick one out of five investment opportunities, labeled Investment 1 through Investment 5, whose return depends on the period-12 price of Security A. Just as in Study 1(a), we ensure incentive compatibility under a wide set of preferences by using only two possible payments per choice problem—receive a bonus of €20 versus not.

Participants are told that Investment 1 "makes a gain" (in effect pays the bonus, see below) iff the selling price of Security A is between 0 and 6,400. Investment 2 makes a gain iff the selling price of Security A is between 6,400 and 12,800, Investment 3 makes a gain iff the selling price is between 12,800 and 19,200, Investment 4 makes a gain iff the selling price is between 19,200 and 25,600, and Investment 5 makes a gain iff the selling price is above 25,600.

Once a participant has chosen her investment the computer simulates the period-12 price of Security A, with a separate simulation for each participant. If the chosen investment makes a gain, the participant receives the bonus. Under any belief the decision maker should, evidently, choose the interval that has the largest probability of containing the period-12 price. Due to the price distribution's large skew, the rational prediction is to choose Investment 1, whose corresponding interval [0 ; 6,400) contains the period-12 price with 80% chance. The other intervals thus have a far lower success chance and the monetary incentive for a participant to choose one of them is far lower. An EGB decision maker, however, perceives a symmetric price distribution $\tilde{Y}_T$ with a mode at 16,000. This implies that the interval containing 16,000 has the highest perceived chance of yielding the bonus. An EGB decision maker therefore chooses Investment 3.

A.1.1.2 Results

The numbers of participants (and percentages in parentheses) choosing Investments 1 through 5 are: $\{Inv.\ 1 : 4\ (6\%), Inv.\ 2 : 19\ (28\%), Inv.\ 3 : 30\ (43\%), Inv.\ 4 : 12\ (17\%), Inv.\ 5 : 4\ (6\%)\}$. The distribution is significantly different from uniform choice (p-value<0.001, chi-square test) and indicates no tendency to choose a mode near zero. Overall, with 94% of the participants significantly more than half of the sample overestimate the true mode (p-value<0.001, one-sided binomial test). While only 6% make the optimal choice of Investment 1, 43% conform with the EGB model and choose Investment 3. The participants give up significant amounts of money due to the bias: while the optimal choice would earn €16.12 in expectation, the observed choice distribution on average earns only €2.07 in expectation per participant.

A.1.2 Treatment Variation HELP of Study 1(b)

A.1.2.1 Treatment description and exponential growth bias prediction

Participants in Study 1(b) are randomly assigned to one of two treatments. Treatment NO_HELP (N=68), which is described in the main text above, presents only the basic explanation. In treatment HELP (N=60) we provide the participants with an additional explanation, leaving the remainder of the instructions unchanged. The additional text (about one written page) gives an explicit calculation of the distribution of compound price changes after two periods. It also points out the asymmetry in the selling price distribution and lists the implicit probabilities of receiving the bonus from choosing Security A for each value of t_A. None of the explanations in HELP adds any substantive information relative to the descriptions in NO_HELP. The only difference is that the relevant distributions are explicit in HELP and implicit in NO_HELP. Any difference in responses under the two conditions must stem from differences in the understanding of these implied truths. Thus, in treatment HELP the EGB cannot influence the subjective beliefs without contradicting the available explanations. We therefore expect the misperception to disappear, i.e., $q_{0.5,i}^{HELP} = 500$.

A.1.2.2 Results

	Share of participants switching from A to B				
Range of subjective median for Security A	Round 1	Round 2	Round 3	Round 4	Round 5
[0 ; 100)	0.000	0.032	0.047	0.046	0.092
[100 ; 500)	0.000	0.016	0.000	0.046	0.046
[500 ; 2,000)	0.703	0.612	0.666	0.676	0.661
[2,000 ; 6,000)	0.109	0.145	0.095	0.138	0.046
[6,000 ; 9,000)	0.063	0.048	0.063	0.046	0.061
[9,000 ; 12,000)	0.063	0.064	0.063	0.000	0.030
[12,000 ; 20,000)	0.031	0.064	0.031	0.462	0.046
[20,000 ; 35,000)	0.016	0.000	0.015	0.000	0.015
[35,000 ; 90,000)	0.000	0.000	0.015	0.000	0.000
[90,000 ; 250,000)	0.000	0.000	0.000	0.000	0.000
[250,000 ; ∞)	0.016	0.016	0.000	0.000	0.000

Table A.1: Subjective medians in HELP for rounds 1-5.

Table A.1 shows that in HELP 70% of responses are at the optimal switching point of Task 3 already in Round 1. Parametric t-tests as well as non-parametric Wilcoxon rank-sum tests confirm that all round-by-round treatment comparisons between HELP and NO_HELP are statistically significant at p-values<0.001. Performance is poor under the NO_HELP condition and much better in HELP.

A.1.3 Low Volatility Treatment Variations of Study 2

A.1.3.1 Experimental Design and Exponential Growth Bias Prediction

In two further treatments of Study 2 (referring to Footnote 13 in Subsection 1.4.1), *Low Volatility Short* (LVS, N=29) and *Low Volatility Long* (LVL, N=29), with $\mu^{h,LVS} = \mu^{h,LVL} = 1.012$ and $\mu^{l,LVS} = \mu^{l,LVL} = 1.011$, the price motion is approximately deterministic (i.e., the price volatility is very low) and the price has positive growth with certainty. The number of time periods until maturity is analogous to HVS and HVL at $T^{LVS} = 14$ and $T^{LVL} = 140$. An additional treatment *Low Volatility Long NoCalculator* (LVL_NC, N=32) is identical to LVL but does not grant the participants access to calculators whereas a hand-held calculator is supplied in the other four treatments in Study 2.[80]

Again, we use the EGB decision maker, who perceives the distribution of absolute returns as constant over time, in order to derive our predictions. With a perceived constant distribution of absolute changes $\tilde{\eta}_t \in \{1.2; 1.1\}$, the EGB decision maker perceives a symmetric distribution (which is correct) of the period-14 (for LVS) and period-140 (for LVL) selling price with mode, mean, and median at $\mathbb{E}(\tilde{Y}_T)$ equal to 116.10 and 261.00, respectively. The EGB model thus predicts that participants underestimate the true median of LVL (495.69) and are quite close to the true value of LVS (117.36).

The elicitation procedure was analogous to that of the other treatments in Study 2.[81] Table A.2 shows the relevant treatment-specific thresholds t_A and Figure A.1 to A.3 illustrate the results.

	Values of t_A *LVS*	Values of t_A *LVL*	Values of t_A *LVL_NC*
Task 1	104.0	185	185
Task 2	104.5	210	210
Task 3	105.5	240	240
Task 4	107.0	290	290
Task 5	109.0	340	340
Task 6	111.5	400	400
Task 7	114.5	460	460
Task 8	118.0	520	520
Task 9	122.0	625	625
Task 10	126.5	850	850

Table A.2: The thresholds t_A by treatment condition.

[80] We included treatment LVL_NC one year after the other treatments. A behavioral difference between treatments LVL and LVL_NC would indicate that the degree to which participants are able to do numerical calculations is a driver of choice.

[81] In contrast to the HVS and HVL treatments, the quantile elicitation is always payoff relevant as the low volatility treatments skip the profit probability elicitation.

A.1.3.2 Results

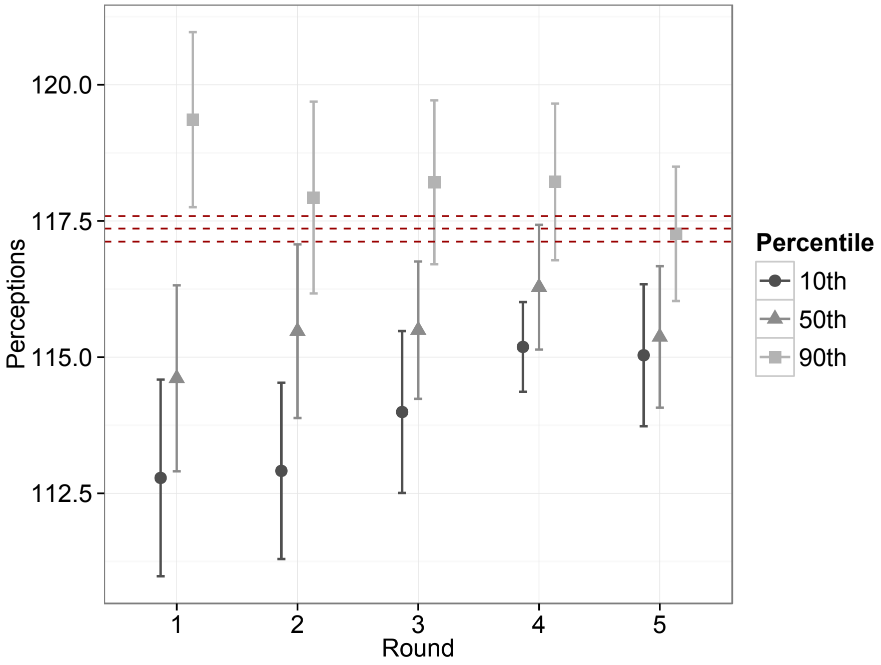


Figure A.1: Point estimates of the participants' subjective quantiles of Security A's selling price distribution in LVS enclosed by 95% confidence intervals. For each of the five rounds, separate estimates refer to the subjective 10th percentiles (circle), subjective medians (triangle), and subjective 90th percentiles (square). Dashed lines indicate rational benchmarks for the 10th percentile (lowest), median (middle) and 90th percentile (uppermost).

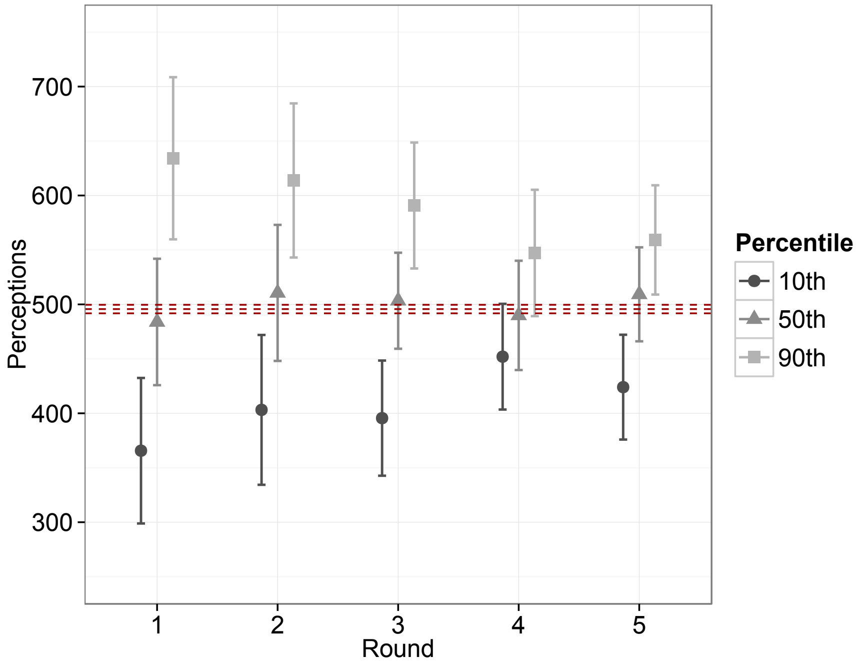


Figure A.2: Point estimates of the participants' subjective quantiles of Security A's selling price distribution in LVL enclosed by 95% confidence intervals. For each of the five rounds, separate estimates refer to the subjective 10th percentiles (circle), subjective medians (triangle), and subjective 90th percentiles (square). Dashed lines indicate rational benchmarks for the 10th percentile (lowest), median (middle) and 90th percentile (uppermost).

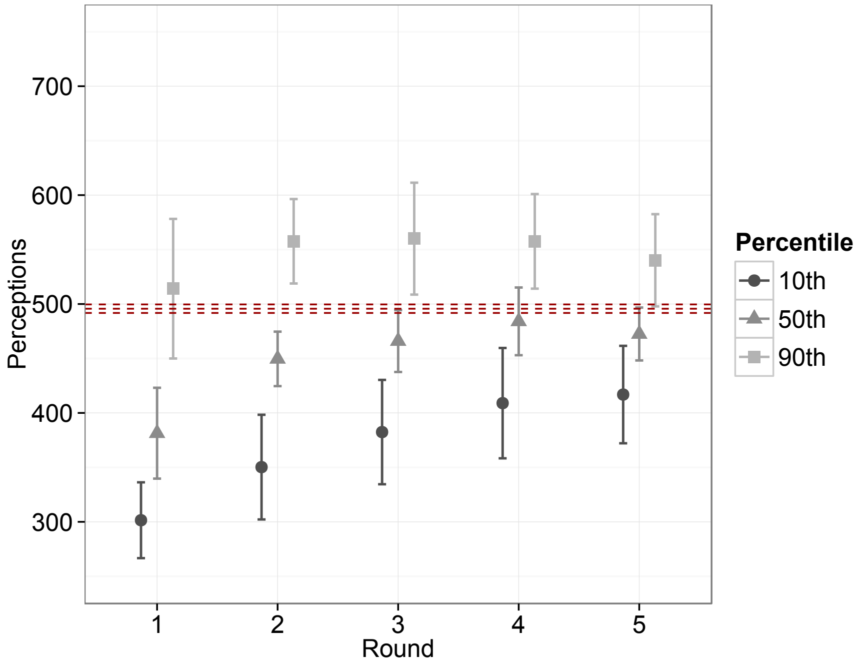


Figure A.3: Point estimates of the participants' subjective quantiles of Security A's selling price distribution in LVL_NC enclosed by 95% confidence intervals. For each of the five rounds, separate estimates refer to the subjective 10th percentiles (circle), subjective medians (triangle), and subjective 90th percentiles (square). Dashed lines indicate rational benchmarks for the 10th percentile (lowest), median (middle) and 90th percentile (uppermost).

A.1.4 Interval Regression Results (Tables)

	10th Percentile		50th Percentile		90th Percentile	
	Mean	Std.Err.	Mean	Std.Err.	Mean	Std.Err.
Rational Model	39.69		88.64		197.98	
EGB Model	35.90		116.10		196.30	
Round 1	83.56	(11.8)	116.56	(12.7)	221.40	(13.5)
Round 2	70.63	(11.8)	108.84	(11.7)	206.84	(11.4)
Round 3	68.74	(11.7)	108.97	(9.5)	191.24	(11.2)
Round 4	62.72	(6.7)	122.41	(9.8)	183.57	(9.2)
Round 5	54.55	(6.3)	120.79	(11.2)	181.92	(11.9)

Table A.3: Interval regression estimates for the mean perceptions of the three elicited percentiles in HVS complemented by EGB and rational predictions at the top of each column.

	10th Percentile		50th Percentile		90th Percentile	
	Mean	Std.Err.	Mean	Std.Err.	Mean	Std.Err.
Rational Model	1.20		29.96		745.58	
EGB Model	0.00		261.00		581.80	
Round 1	82.27	(19.2)	229.80	(27.3)	597.49	(53.2)
Round 2	100.65	(28.9)	202.24	(38.9)	442.75	(50.8)
Round 3	53.68	(9.3)	150.07	(18.2)	373.06	(37.4)
Round 4	79.75	(22.9)	175.52	(24.4)	385.44	(40.9)
Round 5	68.90	(23.4)	165.97	(24.8)	420.60	(59.9)

Table A.4: Interval regression estimates for the mean perceptions of the three elicited percentiles in HVL complemented by EGB and rational predictions at the top of each column.

	10th Percentile		50th Percentile		90th Percentile	
	Mean	Std.Err.	Mean	Std.Err.	Mean	Std.Err.
Rational Model	117.12		117.36		117.59	
EGB Model	115.90		116.10		116.30	
Round 1	112.78	(0.92)	114.61	(0.87)	119.36	(0.82)
Round 2	112.91	(0.83)	115.47	(0.81)	117.93	(0.90)
Round 3	113.99	(0.76)	115.49	(0.64)	118.21	(0.77)
Round 4	115.18	(0.42)	116.28	(0.58)	118.21	(0.73)
Round 5	115.03	(0.67)	115.37	(0.66)	117.26	(0.63)

Table A.5: Interval regression estimates for the mean perceptions of the three elicited percentiles in LVS complemented by EGB and rational predictions at the top of each column.

	10th Percentile		50th Percentile		90th Percentile	
	Mean	Std.Err.	Mean	Std.Err.	Mean	Std.Err.
Rational Model	491.79		495.69		499.63	
EGB Model	260.20		261.00		261.80	
Round 1	365.65	(34.9)	483.87	(29.6)	634.18	(37.9)
Round 2	403.14	(35.1)	510.56	(31.8)	613.80	(36.1)
Round 3	395.55	(26.9)	503.36	(22.4)	590.81	(29.4)
Round 4	451.99	(24.7)	489.91	(25.6)	547.23	(29.5)
Round 5	424.01	(24.5)	509.23	(22.0)	559.20	(25.6)

Table A.6: Interval regression estimates for the mean perceptions of the three elicited percentiles in LVL complemented by EGB and rational predictions at the top of each column.

	10th Percentile		50th Percentile		90th Percentile	
	Mean	Std.Err.	Mean	Std.Err.	Mean	Std.Err.
Rational Model	491.79		495.69		499.63	
EGB Model	260.20		261.00		261.80	
Round 1	301.48	(17.8)	381.39	(21.28)	514.08	(32.7)
Round 2	350.25	(24.5)	449.63	(12.76)	557.62	(19.7)
Round 3	382.42	(24.4)	466.01	(14.51)	560.03	(26.2)
Round 4	408.99	(25.8)	484.08	(15.86)	557.55	(22.1)
Round 5	416.87	(22.8)	472.47	(12.39)	540.17	(21.6)

Table A.7: Interval regression estimates for the mean perceptions of the three elicited percentiles in LVL_NC complemented by EGB and rational predictions at the top of each column.

	10th Percentile		50th Percentile		90th Percentile	
	Sim.Value	Bootstr.Conf.Inter.	Sim.Value	Bootstr.Conf.Inter.	Sim.Value	Bootstr.Conf.Inter.
Rational Model	98.80	[94.95 ; 101.25]	152.98	[146.18 ; 162.40]	310.98	[292.75 ; 322.40]
EGB Model	109.36	[104.21 ; 113.83]	162.72	[156.63 ; 169.16]	227.10	[221.59 ; 230.66]
	Mean	Std.Err.	Mean	Std.Err.	Mean	Std.Err.
Round 1	168.02	(24.2)	316.25	(30.8)	722.22	(95.5)
Round 2	159.65	(17.2)	265.84	(27.7)	513.36	(64.8)
Round 3	151.62	(14.4)	246.15	(21.5)	468.22	(58.2)
Round 4	160.56	(13.5)	315.93	(50.9)	551.12	(75.1)
Round 5	155.61	(14.5)	233.18	(21.7)	477.73	(63.9)

Table A.8: Interval regression estimates for the mean perceptions of the three elicited percentiles in ETF_1 complemented by EGB and rational predictions at the top of each column.

	10th Percentile		50th Percentile		90th Percentile	
	Sim.Value	Bootstr.Conf.Inter.	Sim.Value	Bootstr.Conf.Inter.	Sim.Value	Bootstr.Conf.Inter.
Rational Model	32.39	[29.24 ; 36.31]	152.14	[123.17 ; 178.88]	1,359.42	[1,163.44 ; 1,555.64]
EGB Model	128.09	[112.64 ; 141.51]	288.17	[269.91 ; 307.49]	481.30	[464.79 ; 491.98]
	Mean	Std.Err.	Mean	Std.Err.	Mean	Std.Err.
Round 1	134.87	(21.3)	297.83	(55.7)	582.54	(82.7)
Round 2	156.07	(27.8)	302.48	(43.3)	681.32	(93.8)
Round 3	172.03	(22.7)	295.44	(34.4)	605.39	(86.4)
Round 4	155.59	(29.6)	268.41	(35.3)	636.71	(87.7)
Round 5	145.72	(29.7)	248.39	(41.5)	624.07	(93.6)

Table A.9: Interval regression estimates for the mean perceptions of the three elicited percentiles in ETF_3 complemented by EGB and rational predictions at the top of each column.

A.1.5 Instructions of Treatment ETF_3

Procedure and payment structure

There are five rounds in this experiment. In each round, you have the opportunity to earn a bonus of £5. In what follows, the term *bonus* will always refer to these £5. All bonuses that you earn in any of the five rounds will be paid to you in cash after the experiment.

Each round consists of a number of tasks. In every round, one of these tasks will be chosen by a random draw made on the computer. This task will be paid out for real. That is, if you were successful in the task that the computer picked, you will earn the bonus of £5. If you were unsuccessful in the task that the computer picked, you will not receive a bonus in this round.

The tasks and the structure of the rounds are described on the next pages.

Structure of the rounds

You are an investment manager and have to decide between pairs of risky securities. There are three different investment situations. The three investment situations differ only in the selectable securities. In situation 1, you can either invest in an ETF security or invest in security B1. In situation 2, you can either invest in an ETF security or invest in security B2. And in situation 3, you can either invest in an ETF security or invest in security B3. In all situations, either security, if bought, has to be held for 2000 trading days (roughly 8 years). After the 2000 trading days you sell the security. Depending on your investment success, you have the chance to earn a bonus.

The ETF security has the following properties:

The ETF security can be bought at a price of £100. It is a so-called triple leveraged ETF on the DAX30. The DAX30 is a stock index composed of stocks of 30 major German companies from across all sectors of the economy. This includes automotive companies such as VW and Daimler, utilities such as E.on or banks and insurances such as Deutsche Bank or Allianz. The DAX30 summarizes the stock performance of these 30 companies. Now, the triple leveraged ETF on the DAX30 is a financial product (an *exchange traded fund* or ETF) that moves in the same direction as the DAX30 itself. If the DAX30 increases on a given trading day, then so does the triple leveraged ETF, and vice versa. But the movements of the triple leveraged ETF are stronger than that of the underlying DAX30. More precisely, they are 3 times as strong (hence the name triple leveraged). That is, the triple leveraged ETF gains 3% in value on a day where the DAX30 gains 1% in value and it loses 3% on a day where the DAX30 loses 1%.
For example, suppose on the first trading day the DAX30 increases by 1.5%. Then, the value of your holdings in the triple leveraged ETF increases by 4.5%. That is, the price of the ETF security would be £104.5 after trading day 1. Likewise, if the DAX30 decreases by 1.5% on trading day 1, then the price of the ETF security would be £95.5 after trading day 1. If then on the second trading day the DAX30 decreases by, say, 0.8% the value of your holdings in the ETF security after trading day 2 are 2.4% lower than after trading day 1 (and analogously for gains on trading day 2). And so on, until you sell the ETF security at its price at the end of trading day 2000.
To determine your payment (to be explained in detail later), we will use real-world historical data of the DAX30, going back to the year 1964 when the DAX30 was first compiled. In this period, the magnitudes of daily changes can be summarized as follows: In 90% of all trading days the DAX30 exhibited percentage changes in a range between -1.8% and 1.8%. The average percentage change over all trading days of the DAX30 since 1964 is 0.03%. The overall daily percentage changes of the DAX30 are illustrated in Figure A.4:

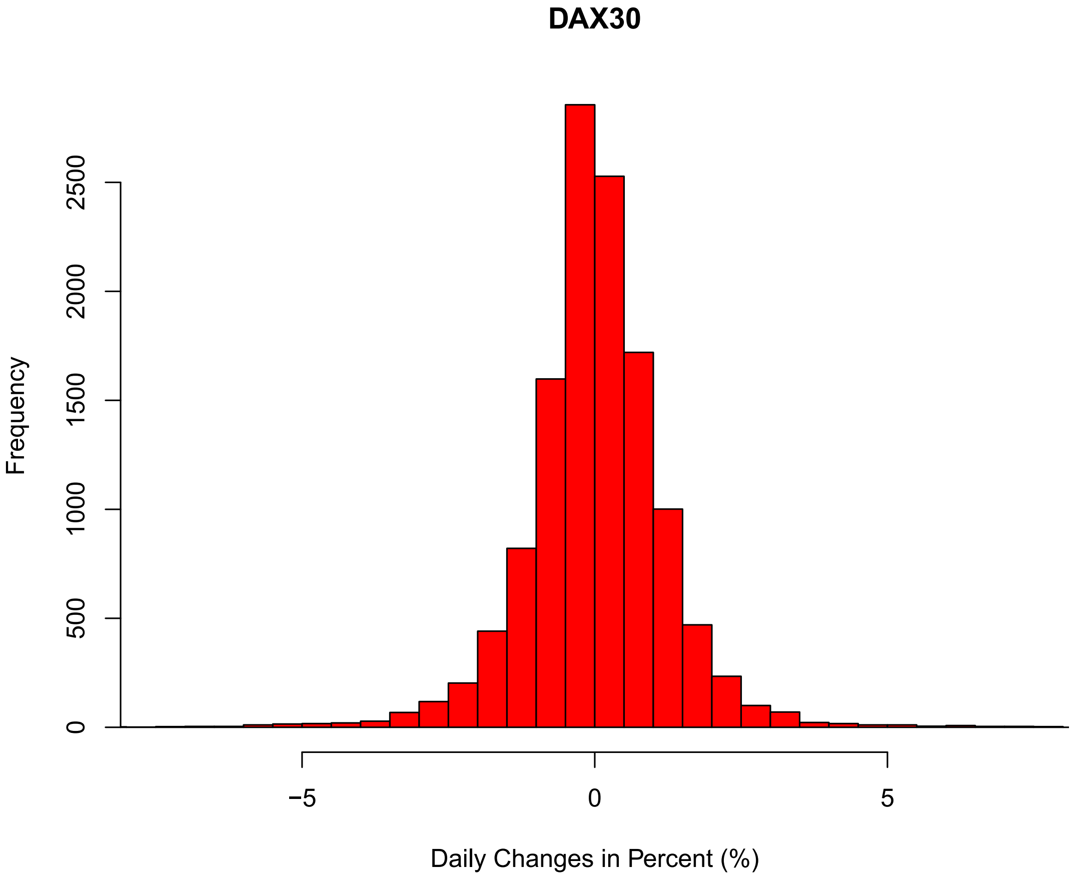


Figure A.4: Daily percentage changes in the total history of the DAX30.

If you choose to invest in the ETF security in the choice task of this experiment, you receive the bonus if the selling price of the ETF security is higher than a threshold. This is explained in detail in the section "Thresholds" below.

The three securities B1, B2, and B3 differ only in the probabilities of giving you the bonus:

Security B1 can be bought at a price of £100. At the end of trading day 2000, you receive the bonus with a probability of 90%. With the remaining 10% you do not receive the bonus.

Security B2 can be bought at a price of £100. At the end of trading day 2000, you receive the bonus with a probability of 50%. With the remaining 50% you do not receive the bonus.

Security B3 can be bought at a price of £100. At the end of trading day 2000, you receive the bonus with a probability of 10%. With the remaining 90% you do not receive the bonus.

Thus, choosing securities B1, B2, or B3, gives you the bonus with probabilities 90%, 50%, and 10%, respectively.
Thresholds

There are 11 different thresholds. The thresholds vary between £30 and £1600. For each of the 11 thresholds you have to choose between investing in the ETF security and investing in one of the securities B1, B2, or B3, respectively. That is, first you have to choose between the ETF security and security B1 for all 11 thresholds, then between the ETF security and security B2 for all 11 thresholds, and finally between the ETF security and security B3 for all 11 thresholds. If, for a given threshold value, you chose the ETF security, you receive the bonus if the final price of the ETF security exceeds that threshold. The choice tasks are listed in tables A.10 to A.12 below.

	Thresholds for the ETF security	Your decision (ETF or B1)
Task 1	30	
Task 2	60	
Task 3	90	
Task 4	140	
Task 5	200	
Task 6	260	
Task 7	330	
Task 8	450	
Task 9	650	
Task 10	1000	
Task 11	1600	

Table A.10: The 11 binary decisions between ETF and B1.

	Thresholds for the ETF security	Your decision (ETF or B2)
Task 1	30	
Task 2	60	
Task 3	90	
Task 4	140	
Task 5	200	
Task 6	260	
Task 7	330	
Task 8	450	
Task 9	650	
Task 10	1000	
Task 11	1600	

Table A.11: The 11 binary decisions between ETF and B2.

	Thresholds for the ETF security	Your decision (ETF or B3)
Task 1	30	
Task 2	60	
Task 3	90	
Task 4	140	
Task 5	200	
Task 6	260	
Task 7	330	
Task 8	450	
Task 9	650	
Task 10	1000	
Task 11	1600	

Table A.12: The 11 binary decisions between ETF and B3.

The following examples are intended to make you more familiar with the mechanics of these tasks:

Example 1:
Consider *Task 1* in situation 1 (choice between the ETF security and security B1, Table A.10). The threshold for the ETF security is £30. Suppose that you decide to buy the ETF security. If the selling price of the ETF security is higher than £30, you receive the bonus. If the selling price is less than or equal to £30, you do not receive the bonus. Now, suppose instead that you decide to buy security B1. This means that you choose a lottery that pays the bonus with a 90% probability.

Example 2:
Now consider *Task 2* in situation 1, where the threshold of the ETF security is higher than in the previous example, at £60, and security B1 again yields the bonus with a probability of 90%. Suppose that you decide to buy the ETF security. In this case, if the ETF security's selling price is higher than £60, you receive the bonus. Otherwise, you do not receive the bonus. If instead you decide to buy security B1, you again receive the bonus with 90% probability.

And so on, analogously for *Task 3*, *Task 4*, etc., until *Task 11*. You then repeat this process for situations 2 and 3. However, there are two restrictions to your choices.

Restriction 1: Within 11 tasks, you may only switch once from choosing the ETF security to choosing the B type security. That is, if you choose to invest in the ETF security for the first few tasks and then switch to the B type security, you have to stay with this choice until *Task 11*. Note that this restriction cannot diminish your estimated chances of receiving the bonus. The thresholds for the ETF security increase with each task. The higher the threshold the less likely it is that the security's selling price will exceed it. Thus, if you believe that the ETF security in a given task gives you a larger chance of receiving the bonus than the B type security, then you should choose the ETF security for all lower tasks as well. Hence, you should not switch back and forth between the ETF security and the B type security.

Restriction 2: Within one round, the threshold at which you switch from the ETF security to security B1 cannot exceed the threshold at which you switch from the ETF security to security B2 which cannot exceed the threshold at which you switch from the ETF security to security B3.
Again, this restriction cannot diminish your estimated chances of receiving the bonus. To see this, suppose that for a given task you prefer security B2 to the ETF security. This implies that you estimate the chance of the ETF security's selling price being above that threshold to be less than 50%. But then you should also prefer security B1 to the ETF security at this threshold, as security B1's odds of returning the bonus are even better than those of security B2. In the comparison between the ETF security and security B2, you should therefore choose the ETF security for at least as many threshold values as in the comparison between the ETF security and security B1. That is, you should switch to security B1 no later then when you switch to security B2. The same logic applies when comparing switching in situations 2 and 3.
In order to help you not violating this restriction, for situations 2 and 3 during the whole experiment, the number of tasks at which you chose the ETF security in the previous situation is shown to you on the right side of the computer screen. This shown number is the minimal number of tasks at which you have to choose the ETF security in the current situation.

When you have completed the choice tasks we will ask you to state your believed probability of the ETF security making a profit, i.e., whether the final price of the ETF security lies above 100. The details are outlined in the following paragraphs.

Stating probabilities

In order to elicit your believed probability of the ETF security making a profit in the 2000 trading days, we will again ask you to compare an investment in the ETF security with alternative investments.
More specifically, we want you to consider a pool of different securities of which you can choose one at random. Imagine that you pick one of these securities from a bag of securities, without being able to see inside the bag. After you have picked the security, you can see it and observe its probability of making a profit. Each one of these securities has some probability of making a profit after the 2000 trading days. These probabilities can assume any decimal number between 0% and 100%, and each number has equal probability. For example, there is one alternative security that makes a profit with a probability of 20.5%. If you invest in this security you receive the bonus with a probability of 20.5%. You can either invest in the ETF security or in the randomly chosen alternative security. The general payoff rule is that for any security that you invest in, you receive the bonus if the security makes a profit.

To determine your investment decision between the ETF security and the randomly chosen security, we would like you to state an integer number X that satisfies the following statement:

"If I choose an alternative security at random with a profit probability less or equal to X, I keep investing in the ETF security. However, if an alternative security is chosen which has a profit probability greater than X, I invest in this alternative security."

The number X that we want you to write down is therefore equal to your believed probability of the ETF security making a profit. Just in case you doubt that it is indeed in your best interest to answer truthfully, you can go through the following explanation (smaller print):

Suppose you believe that the profit probability of the ETF security is 50%. Then you should state 50% as the threshold. Why? Assume you lie and state a higher number than you really believe, say, 60%. Then one of three things will happen. First it could happen that the drawn security has a profit probability which is higher than your believed number but lower than your stated number, e.g., 59%. You would then still, according to the rules, advise to invest in the ETF security which in your opinion has a profit probability of 50%. If you had stated your true belief, however, you would advise to invest in the alternative security. Thus, your chances to receive the bonus would be higher if you state you true belief. The two remaining cases are draws from the alternative security pool that are securities with probabilities either greater than your stated 60% or lower than your believed 50%. In both cases you cannot benefit from lying; if your draw has a probability greater than 60% you would have prefered it either way. Conversely, a draw with less than 50% probability always lets you stay with the ETF security.
Now suppose you state a number which is lower than your believed 50%, say, 40%. Then again, it could happen that the chosen security has a winning probability which is in between your believed 50% and your stated 40%, say, 45%. You would then advise to invest in this security, which gives you the bonus with a 45% chance. If you had stated your true belief, you would have advised to invest in the ETF security, which would have given you a winning probability of 50% instead. So misstating your belief made you worse off. Finally, if you draw an security with a winning probabilty of less than your stated 40% or above your believed 50% your odds of receiving the bonus are the same as when you state your true belief.
Summarizing the above, it is generally true that not writing down what you really believe would worsen your chances of getting the bonus.

After you have stated your believed probability of the ETF security making a profit at the end of 2000 trading days, the round is completed. The whole procedure is repeated five times. Thus, every round consists of the decisions between the ETF security and the securities B1, B2, and B3, respectively, and your estimate of your believed profit probability of the ETF security.

Payment and Feedback

For each round, one of the 33 choice tasks (choosing between the ETF security and one of the securities B1, B2, or B3) is selected at random by a computerized draw. Each of the tasks is chosen with the same probability. Then, the computer will be used to make a further draw to choose between the selected task and your final question of the round, where you stated the profit probability of the ETF security. Each of the two numbers is picked with equal probability ("fifty-fifty"). Depending on whether the choice task or your stated profit probability is picked by the computer, you will receive the bonus if you were successful in the selected choice or estimation task. If you were unsuccessful, you will not receive a bonus in this round. This procedure is then conducted for each of the 5 rounds.

As mentioned above, the DAX30 and leveraged ETFs are real-world financial products. Thus, we will use real-world historical data. More precisely, a computer will randomly pick a trading day as day 1 of the 2000 trading days period. This can be any trading day between 31 December, 1964, and 11 November, 2004. All days are chosen with the same probability and the draws are made independently for each round and each participant. The 2000 trading days following this day are your investment period. You receive the bonus if the selling price of the ETF security is above the relevant threshold after this period. Let's look at an example:

After each round, the computer makes a "fifty-fifty" selection between the choice tasks and the probability stating task. Suppose first that it picks the choice tasks. Then, one of the 33 choice tasks is chosen at random (and with equal probability) to be the relevant one. Suppose that in round 1 the computer picks *Task 3* of the comparison between the ETF security and security B1 to be the payoff relevant task. Now, depending on your choice in this task you receive the bonus for round 1 or not. That is, if you chose the ETF security in situation 1, *Task 3*, and the ETF security's selling price was above the threshold for *Task 3*, you receive the bonus in round 1. If instead you chose security B1 in that task you get the bonus if B1 was successful in that task (which happens with a 90% probability).

Alternatively, suppose that the computer picks the probability stating task to be the relevant task in round 1. Then you get the bonus if you are successful in this task. That is, the computer will pick and simulate a security in accordance with your stated probability (i.e., the integer number X) as described above and return the bonus if this security or the ETF security - depending on your choice - does return a profit.

After each round, you will receive the realized price of the ETF security after 2000 trading days as feedback. This realized price is generated from real-world data by a computer. The computer randomly picks the 2000 trading days period according to the procedure described above.

Are there questions about the tasks or payment rules in this experiment? If so, please raise your hand and we will help you at your desk. If there are no further questions at this point, you will now face a brief understanding test. Only if you answer all questions correctly, you will proceed to the actual tasks.

Calculators are allowed during the whole experiment.

A.2 Investment Behavior in Peer-to-Peer Lending

A.2.1 Default Probability Estimation

A.2.1.1 Discrete Time Hazard Model

We denote D as the random variable indicating the time period of a loan's default, with values $d_1<d_2<d_3<...$, and write the default probabilities as $p(d_j) = Pr\{D = d_j\}$. We then define a loan's survival function at a specific d_j as the probability that the loan didn't default until $D \geq d_j$ resulting in $S(d_j) = Pr\{D \geq d_j\} = \sum_{D \geq d_j} p(d_j)$. Next, we define the discrete time hazard rate, $h(d_j)$, as a loan's conditional probability of defaulting at time d_j given the loan hasn't defaulted up to that point, so that

$$h(d_j) = Pr\{D = d_j | D \geq d_j\} = \frac{p(d_j)}{S(d_j)}. \tag{A.1}$$

To estimate the discrete time hazard rate conditional on our observable loan characteristics and the current time period, we follow Cox (1972) and use the logistic hazard model approach. The model is fitted by running a logistic regression on a set of pseudo observations generated from our original data set. That is, we generate default indicators y_{ij} that are valued 1 if loan i defaulted at time j and zero other wise. This results in a number of default indicators per loan that is equal to the number of actually observed monthly payments of that loan. Each of these generated indicators is merged with the loan specific covariates, $\mathbf{x_i}$, and a consecutively numbered time index j. Since $Pr\{y_{ij} = 1 | D \geq d_j\} = Pr\{D = d_j | D \geq d_j\} = h_i(d_j)$ the hazard rate can be estimated with the help of the usual maximum likelihood procedures in a binary response data case, with the log likelihood equal to:

$$\text{logL} = \sum_{i=1}^{N} \sum_{j=1}^{D} \left[y_{ij} \text{log}(h_i(d_j, X_i)) + (1 - y_{ij}) \text{log}(1 - h_i(d_j, X_i)) \right], \tag{A.2}$$

where the hazard rate is assumed to have a logistic functional form. The loan specific characteristics such as requested loan amount, offered loan rate, loan duration, loan purpose as well as the borrower characteristics age, gender, schufa rating, financial burden, employment status and place of residence are captured by X_i. Using the predicted values for $\hat{h}_i(d_j, X_i)$ we are able to calculate the survival function

$$\hat{S}(d_j, X_i) = \prod_{j=1}^{d_j - 1} (1 - \hat{h}_i(d_j, X_i)) \tag{A.3}$$

and with the help of Equation (A.1) we can solve for the default probabilities, $\hat{p}(d_j)$.

A.2.1.2 Estimation Results

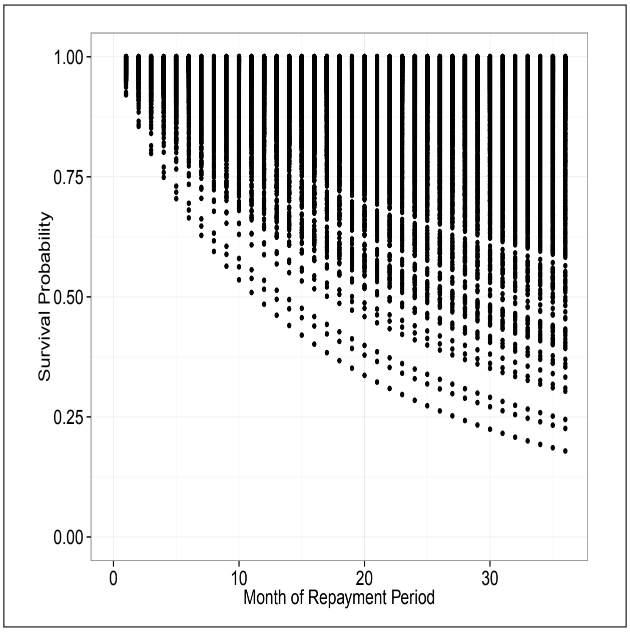


Figure A.5: Evolution paths of the survival probability over time for all loans with a duration of 36 months.

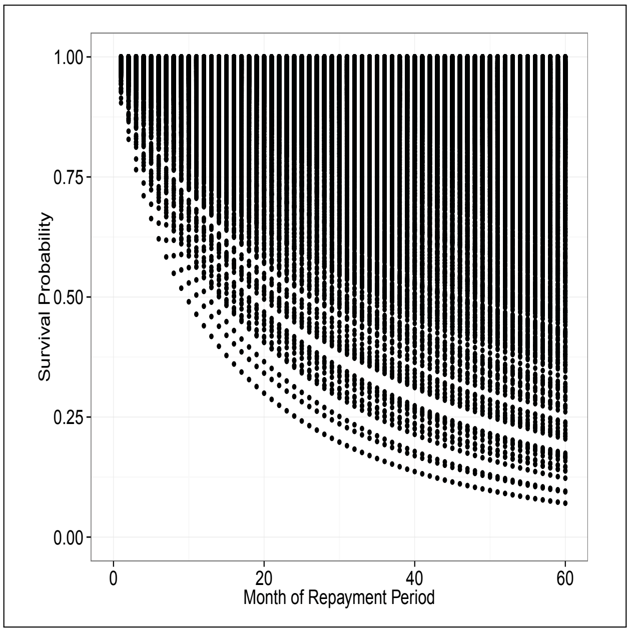


Figure A.6: Evolution paths of the survival probability over time for all loans with a duration of 60 months.

A.2.2 Additional Tables

	SIMEX-Model(1)	SIMEX-Model(2)	SIMEX-Model(3)
E(IRR)	0.104*	0.679***	0.583***
	(0.05)	(0.21)	(0.19)
$borrower_{male}$	-0.018		
	(0.09)		
$borrower_{male}*$ E(IRR)	0.502***		
	(0.06)		
$borrower_{age}$		-0.016***	
		(0.00)	
$borrower_{age}*$ E(IRR)		-0.010***	
		(0.00)	
$borrower_{north}$			0.010
			(0.09)
$borrower_{north}*$ E(IRR)			-0.181***
			(0.06)
Fixed Effects	No	No	No

Significance levels: $*$: <10% $**$: $<$ 5% $***$: $<$ 1%

Table A.13: Ordinal logit results with correction for measurement error via the SIMEX method.

Dependent Variable: $share\_funded$		
Explanatory Variable	Model (3)	SIMEX-Model(3)
$loan\_size$	-0.031***	-0.031***
	(0.02)	(0.02)
$job1$	0.087	0.100
	(0.22)	(0.31)
$job2$	0.062	0.005
	(0.32)	(0.28)
$job3$	0.061	-0.049
	(0.25)	(0.21)
$job4$	-0.166	-0.079
	(0.30)	(0.22)
$job5$	-0.050	0.024
	(0.23)	(0.15)
$schufaB$	-0.683***	-0.594***
	(0.18)	(0.15)
$schufaC$	-0.245	0.074
	(0.22)	(0.25)
$schufaD$	-1.196***	-0.898***
	(0.22)	(0.26)
$schufaE$	-1.244***	-0.861***
	(0.23)	(0.27)
$schufaF$	-1.706***	-1.107***
	(0.24)	(0.35)
$schufaG$	-2.236***	-1.291***
	(0.30)	(0.34)
$schufaH$	-2.640***	-1.305***
	(0.39)	(0.41)
$kdf2$	-0.518***	-0.648***
	(0.15)	(0.41)
$kdf3$	-1.211***	-1.358***
	(0.15)	(0.17)
$kdf4$	-2.188***	-2.417***
	(0.16)	(0.19)
$loan\_rate$	0.178***	-0.006***
	(0.04)	(0.02)
$loan\_duration\_long$	-0.076***	-0.970***
	(0.12)	(0.16)
$description\_length$	0.000***	0.000***
	(0.00)	(0.00)
$business\_loan$	0.388***	0.287***
	(0.12)	(0.11)
fee_new	0.274**	0.273**
	(0.13)	(0.14)
κ_1	-3.298	-4.839
	(0.62)	(0.84)
κ_2	-2.604	-4.140
	(0.62)	(0.83)

Significance levels: $*$: <10% $**$: $<$ 5% $***$: $<$ 1%

Table A.14: Fixed effects.

A.3 Binary Choice Belief Elicitation

A.3.1 Incentive Compatibility

A potential weakness of my adaptive approach is that if subjects knew how the questions are selected based on their previous answers, they could have an incentive to make false statements about their beliefs in order to increase their experimental payoffs by “gaming the system.” That is, when asked to make a choice between two events a subject could lie in order to get an easier next question, which would pay out with a higher likelihood.

For example, taking a look at Figure 3.2 one can see that for the first set of questions after each answer the chosen interval decreases for the next question while the rejected one becomes larger. A subject answering truthfully is confronted with ever more difficult questions. Thus, a subject could lie on early questions in order to create future questions that are easier to answer and thereby more likely to be profitable in an incentivized experiment.

It is not clear how likely such a scenario is, if it would benefit strategizing subjects by much (since subjects have to give up potential payoffs of the questions at which they lie to gain payoffs from easier future questions), and if it would alter the elicited parameters by a significant margin. Anyhow, the existence of subjects who not only understand the adaptive approach but who are also able to compute by themselves what next question would be chosen based on their previous answers and additionally use this knowledge to figure out at which questions it would be optimal to lie in order to increase the overall payoff can not be ruled out for certain.

Fortunately, economic theory provides a variety of solutions for situations with adversely strategizing participants in experiments. Wang et al. (2010) describe three of such design remedies, of which I just shortly summarize the one that is most simple and easy to apply to my design.

First, one question q is chosen at random from the set of all possible questions Q_{all} to be payoff relevant after the conclusion of my adaptive questionnaire. Second, there is adaptive sampling of questions resulting in a subset of asked questions Q_{opt} of the set Q_{all}. Third, after answers to all optimal questions are collected and if question q was among the optimally chosen questions its answer is payoff relevant. For the case of $q \in Q_{all} \setminus Q_{opt}$ the subject has to give an additional answer.

Under this incentive scheme subjects still can strategize to get to answer easier next questions. The chances of those questions, however, to be payoff relevant are left unchanged by strategizing now and, thus, my adaptive design is rendered incentive compatible. The only weak point of this scheme is that if Q_{all} is much larger than Q_{opt} then the answers to the optimal questions become unlikely to be payoff relevant, which decreases the incentive to answer truthfully.

A.3.2 Analytical Derivation of the Hessian

An analytical derivation of the Hessian helps to shave some milliseconds off the evaluation time if I feed PyMC with analytical Hessians, instead of PyMC having to (a) approximate the posterior with a normal and (b) get the Hessian off of that. The log-likelihood of a single observation then is

$$\begin{aligned} l_i\left(\mu,\sigma^2,\eta^2\right) = {} & 1_{\{choice_i='left'\}} \\ & \cdot \ln\left(\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right) \\ & + 1_{\{choice_i='right'\}} \\ & \cdot \ln\left(1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right), \end{aligned} \tag{A.4}$$

where $\Phi\left(\,\cdot\,\right)$ is the standard normal cumulative distribution function, $F\left(\,\cdot\,\right)$ is the utility difference, which just boils down to a difference of differences of normal cumulative distribution functions with parameters μ and σ^2, and the noise parameter η^2. Moreover, for numerical reasons I'll rescale η^2 by ρ. The observation's gradient contribution then is

$$\nabla l_i\left(\mu,\sigma^2,\eta^2\right) = \begin{bmatrix} 1_{\{choice_i='left'\}} \left[\frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)} \frac{1}{\sqrt{\rho\eta^2}} \frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu}\right] \\ +1_{\{choice_i='right'\}} \left[-\frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)} \frac{1}{\sqrt{\rho\eta^2}} \frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu}\right] \\ \\ 1_{\{choice_i='left'\}} \left[\frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)} \frac{1}{\sqrt{\rho\eta^2}} \frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}\right] \\ +1_{\{choice_i='right'\}} \left[-\frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)} \frac{1}{\sqrt{\rho\eta^2}} \frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}\right] \\ \\ 1_{\{choice_i='left'\}} \left[-\frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)} \frac{1}{2\sqrt{\rho}\left(\eta^2\right)^{\frac{3}{2}}} F\left(\mu,\sigma^2\right)\right] \\ +1_{\{choice_i='right'\}} \left[\frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)} \frac{1}{2\sqrt{\rho}\left(\eta^2\right)^{\frac{3}{2}}} F\left(\mu,\sigma^2\right)\right] \end{bmatrix}$$

where $\phi\left(\,\cdot\,\right)$ is the standard normal probability density function. The observation's Hessian contribution therefore has the form

$$H = \begin{matrix} \frac{\partial^2 l_i}{\partial\mu^2} & \frac{\partial^2 l_i}{\partial\mu\partial\sigma^2} & \frac{\partial^2 l_i}{\partial\mu\partial\eta^2} \\ \frac{\partial^2 l_i}{\partial\sigma^2\partial\mu} & \frac{\partial^2 l_i}{\partial\sigma^{22}} & \frac{\partial^2 l_i}{\partial\sigma^2\partial\eta^2} \\ \frac{\partial^2 l_i}{\partial\eta^2\partial\mu} & \frac{\partial^2 l_i}{\partial\eta^2\partial\sigma^2} & \frac{\partial^2 l_i}{\partial(\eta^2)^2} \end{matrix}$$

where

$$\begin{aligned} \frac{\partial^2 l_i}{\partial\mu^2} =& 1_{\{choice_i='left'\}} \left[\frac{\left[\phi'\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\Phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right) - \left(\phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(\Phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\right)^2} \right. \\ & \left. \cdot \left(\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu}\right)^2 + \frac{\phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)}{\Phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)}\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial^2 F\left(\mu,\sigma^2\right)}{\partial\mu^2}\right] \\ & + 1_{\{choice_i='right'\}} \left[-\frac{\left[\phi'\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\left(1-\Phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\right) + \left(\phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(1-\Phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\right)^2} \right. \\ & \left. \cdot \left(\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu}\right)^2 - \frac{\phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)}{1-\Phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)}\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial^2 F\left(\mu,\sigma^2\right)}{\partial\mu^2}\right] \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 l_i}{\partial\sigma^2\partial\mu} = \frac{\partial^2 l_i}{\partial\mu\partial\sigma^2} =& 1_{\{choice_i='left'\}} \left[\frac{\left[\phi'\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\Phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right) - \left(\phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(\Phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\right)^2} \right. \\ & \cdot \left(\frac{1}{\sqrt{\rho\eta^2}}\right)^2 \frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu} \\ & \left. + \frac{\phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)}{\Phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)}\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial^2 F\left(\mu,\sigma^2\right)}{\partial\sigma^2\partial\mu}\right] \\ & + 1_{\{choice_i='right'\}} \\ & \cdot \left[-\frac{\left[\phi'\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\left(1-\Phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\right) + \left(\phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(1-\Phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)\right)^2} \right. \\ & \cdot \left(\frac{1}{\sqrt{\rho\eta^2}}\right)^2 \frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu} \\ & \left. - \frac{\phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)}{1-\Phi\left(\frac{F(\mu,\sigma^2)}{\sqrt{\rho\eta^2}}\right)}\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial^2 F\left(\mu,\sigma^2\right)}{\partial\sigma^2\partial\mu}\right] \end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l_i}{\partial\left(\sigma^2\right)^2} =& 1_{\{choice_i='left'\}} \left[\frac{\left[\phi'\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right) - \left(\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2}\right. \\ &\left. \cdot \left(\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}\right)^2 + \frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial^2 F\left(\mu,\sigma^2\right)}{\partial\left(\sigma^2\right)^2}\right] \\ &+ 1_{\{choice_i='right'\}} \\ &\cdot \left[-\frac{\left[\phi'\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\left(1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right) + \left(\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2}\right. \\ &\left. \cdot \left(\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}\right)^2 - \frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial^2 F\left(\mu,\sigma^2\right)}{\partial\left(\sigma^2\right)^2}\right]\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l_i}{\partial\eta^2\partial\mu} = \frac{\partial^2 l_i}{\partial\mu\partial\eta^2} =& 1_{\{choice_i='left'\}} \left[-\frac{1}{2}\frac{1}{\sqrt{\rho}}\frac{1}{\left(\eta^2\right)^{\frac{3}{2}}}\right. \\ &\cdot \left[\frac{\left[\phi'\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right) - \left(\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2}\right. \\ &\left.\left. \cdot \frac{1}{\sqrt{\rho\eta^2}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu}F\left(\mu,\sigma^2\right) + \frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu}\right]\right] \\ &+ 1_{\{choice_i='right'\}} \left[+\frac{1}{2}\frac{1}{\sqrt{\rho}}\frac{1}{\left(\eta^2\right)^{\frac{3}{2}}}\right. \\ &\cdot \left[\frac{\left[\phi'\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\left(1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right) + \left(\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2}\right. \\ &\left.\left. \cdot \frac{1}{\sqrt{\rho\eta^2}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu}F\left(\mu,\sigma^2\right) + \frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu}\right]\right]\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 l_i}{\partial\eta^2\partial\sigma^2} = \frac{\partial^2 l_i}{\partial\sigma^2\partial\eta^2} =& 1_{\{choice_i='left'\}}\Bigg[-\frac{1}{2}\frac{1}{\sqrt{\rho}}\frac{1}{\left(\eta^2\right)^{\frac{3}{2}}} \\
&\cdot\left[\frac{\left[\phi'\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)-\left(\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2}\right. \\
&\left.\cdot\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}F\left(\mu,\sigma^2\right)+\frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}\right]\Bigg] \\
&+1_{\{choice_i='right'\}}\Bigg[+\frac{1}{2}\frac{1}{\sqrt{\rho}}\frac{1}{\left(\eta^2\right)^{\frac{3}{2}}} \\
&\cdot\left[\frac{\left[\phi'\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\left(1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)+\left(\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2}\right. \\
&\left.\cdot\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}F\left(\mu,\sigma^2\right)+\frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}\right]\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 l_i}{\partial\left(\eta^2\right)^2} =& 1_{\{choice_i='left'\}}\left[\frac{\left[\phi'\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)-\left(\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2}\right. \\
&\left.\cdot\left(\frac{1}{2}\frac{1}{\sqrt{\rho}}\frac{1}{\left(\eta^2\right)^{\frac{3}{2}}}F\left(\mu,\sigma^2\right)\right)^2+\frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}\frac{3}{4}\frac{1}{\left(\eta^2\right)^{\frac{5}{2}}}F\left(\mu,\sigma^2\right)\right] \\
&+1_{\{choice_i='right'\}}\left[-\frac{\left[\phi'\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\left(1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)+\left(\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2}\right. \\
&\left.\cdot\left(\frac{1}{2}\frac{1}{\sqrt{\rho}}\frac{1}{\left(\eta^2\right)^{\frac{3}{2}}}F\left(\mu,\sigma^2\right)\right)^2+\frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}\frac{3}{4}\frac{1}{\sqrt{\rho}}\frac{1}{\left(\eta^2\right)^{\frac{5}{2}}}F\left(\mu,\sigma^2\right)\right]
\end{aligned}
$$

Written more concisely:

$$
\begin{aligned}
\frac{\partial^2 l_i}{\partial\mu^2} &= CF \cdot \begin{bmatrix} \left(\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu}\right)^2 \\ \frac{1}{\sqrt{\rho\eta^2}}\frac{\partial^2 F\left(\mu,\sigma^2\right)}{\partial\mu^2} \end{bmatrix} \\
\frac{\partial^2 l_i}{\partial\sigma^2\partial\mu} = \frac{\partial^2 l_i}{\partial\mu\partial\sigma^2} &= CF \cdot \begin{bmatrix} \frac{1}{\eta^2}\frac{1}{\sqrt{\rho}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu} \\ \frac{1}{\sqrt{\rho\eta^2}}\frac{\partial^2 F\left(\mu,\sigma^2\right)}{\partial\sigma^2\partial\mu} \end{bmatrix} \\
\frac{\partial^2 l_i}{\partial\left(\sigma^2\right)^2} &= CF \cdot \begin{bmatrix} \left(\frac{1}{\sqrt{\rho\eta^2}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}\right)^2 \\ \frac{1}{\sqrt{\rho\eta^2}}\frac{\partial^2 F\left(\mu,\sigma^2\right)}{\partial(\sigma^2)^2} \end{bmatrix} \\
\frac{\partial^2 l_i}{\partial\eta^2\partial\mu} = \frac{\partial^2 l_i}{\partial\mu\partial\eta^2} &= CF \cdot \begin{bmatrix} -\frac{1}{2}\frac{1}{\sqrt{\rho}}\frac{1}{(\eta^2)^2}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu}F\left(\mu,\sigma^2\right) \\ -\frac{1}{2}\frac{1}{\sqrt{\rho}}\frac{1}{(\eta^2)^{\frac{3}{2}}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu} \end{bmatrix} \\
\frac{\partial^2 l_i}{\partial\eta^2\partial\sigma^2} = \frac{\partial^2 l_i}{\partial\sigma^2\partial\eta^2} &= CF \cdot \begin{bmatrix} -\frac{1}{2}\frac{1}{\sqrt{\rho}}\frac{1}{(\eta^2)^2}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}F\left(\mu,\sigma^2\right) \\ -\frac{1}{2}\frac{1}{\sqrt{\rho}}\frac{1}{(\eta^2)^{\frac{3}{2}}}\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2} \end{bmatrix} \\
\frac{\partial^2 l_i}{\partial\eta^{22}} &= CF \cdot \begin{bmatrix} \left(\frac{1}{2}\frac{1}{\sqrt{\rho}}\frac{1}{(\eta^2)^{\frac{3}{2}}}F\left(\mu,\sigma^2\right)\right)^2 \\ \frac{3}{4}\frac{1}{\sqrt{\rho}}\frac{1}{(\eta^2)^{\frac{5}{2}}}F\left(\mu,\sigma^2\right) \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\text{with } CF = \Bigg(& 1_{\{choice_i='left'\}} \begin{bmatrix} \frac{\left[\phi'\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)-\left(\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2} \\ \frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)} \end{bmatrix} \\
& +1_{\{choice_i='right'\}} \begin{bmatrix} -\frac{\left[\phi'\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\left(1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)+\left(\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2\right]}{\left(1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)\right)^2} \\ -\frac{\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)}{1-\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)} \end{bmatrix} \Bigg)^{\top} .
\end{aligned}
$$

To implement this numerically, it is best to split this up into a few components and then to assemble the Hessian step by step. All elements of the Hessian are products of the following components:

1. $F\left(\mu, \sigma^2\right)$
2. $\frac{1}{\eta^2}$
3. $\Phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)$
4. $\phi\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)$
5. $\phi'\left(\frac{F\left(\mu,\sigma^2\right)}{\sqrt{\rho\eta^2}}\right)$
6. $\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\mu}$
7. $\frac{\partial F\left(\mu,\sigma^2\right)}{\partial\sigma^2}$
8. $\frac{\partial^2 F\left(\mu,\sigma^2\right)}{\partial\mu^2}$
9. $\frac{\partial^2 F\left(\mu,\sigma^2\right)}{\partial\sigma^2\partial\mu}$
10. $\frac{\partial^2 F\left(\mu,\sigma^2\right)}{\partial\sigma^{2 2}}$.

From these components one can also easily assemble the gradient.

A.3.3 Prior Densities

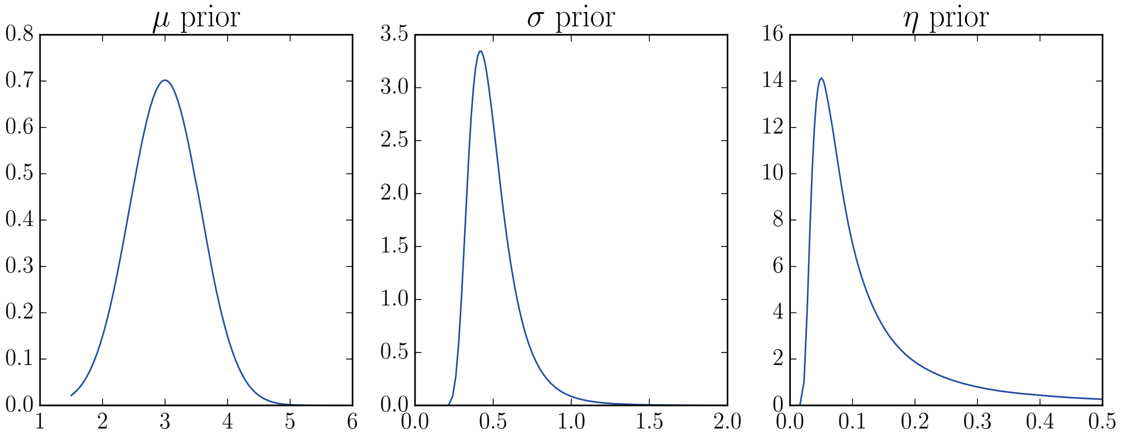


Figure A.7: Prior distributions for μ, σ and η in the Gasoline Price domain.

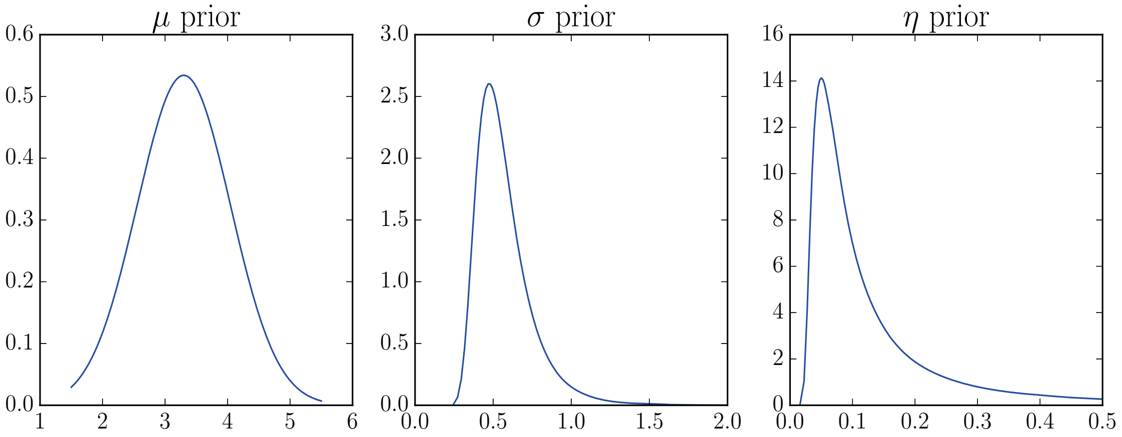


Figure A.8: Prior distributions for μ, σ and η in the Diesel Price domain.

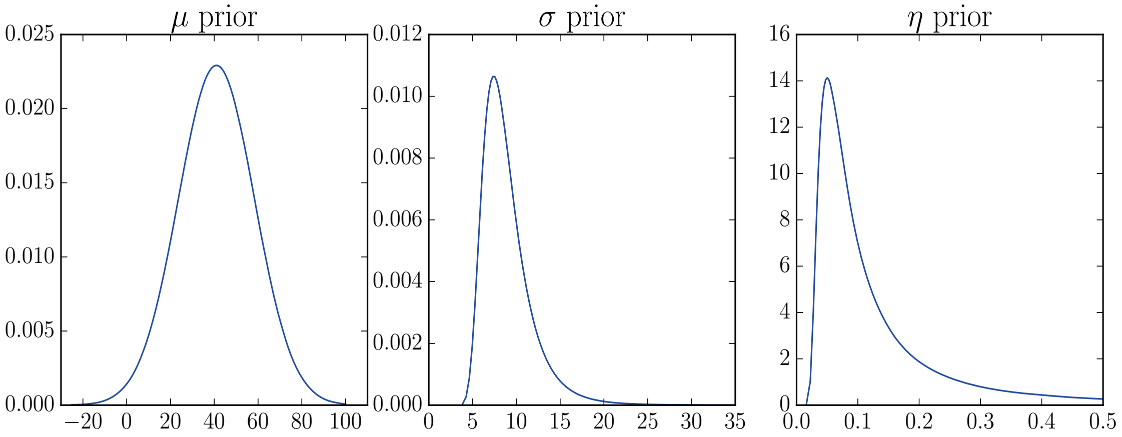
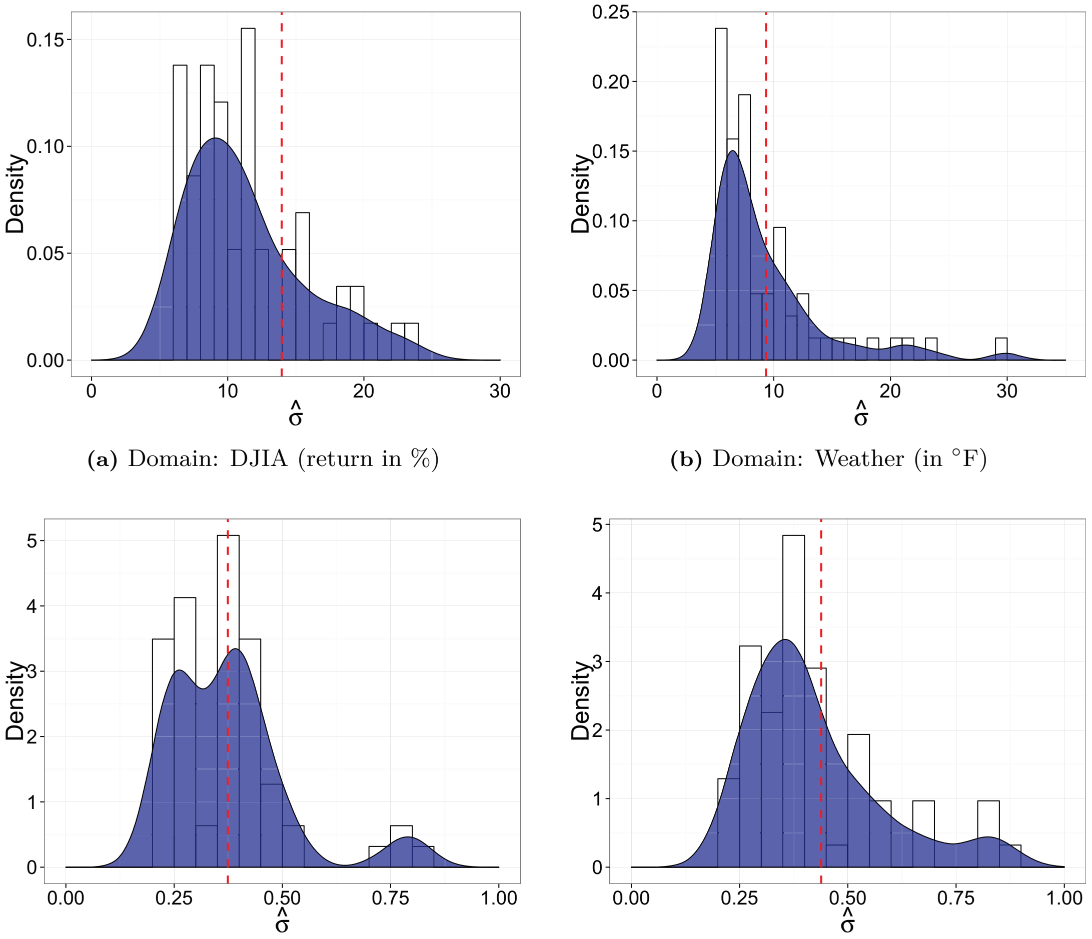


Figure A.9: Prior distributions for μ, σ and η in the Weather domain.

A.3.4 Subject Pool Densities of Parameter Estimates



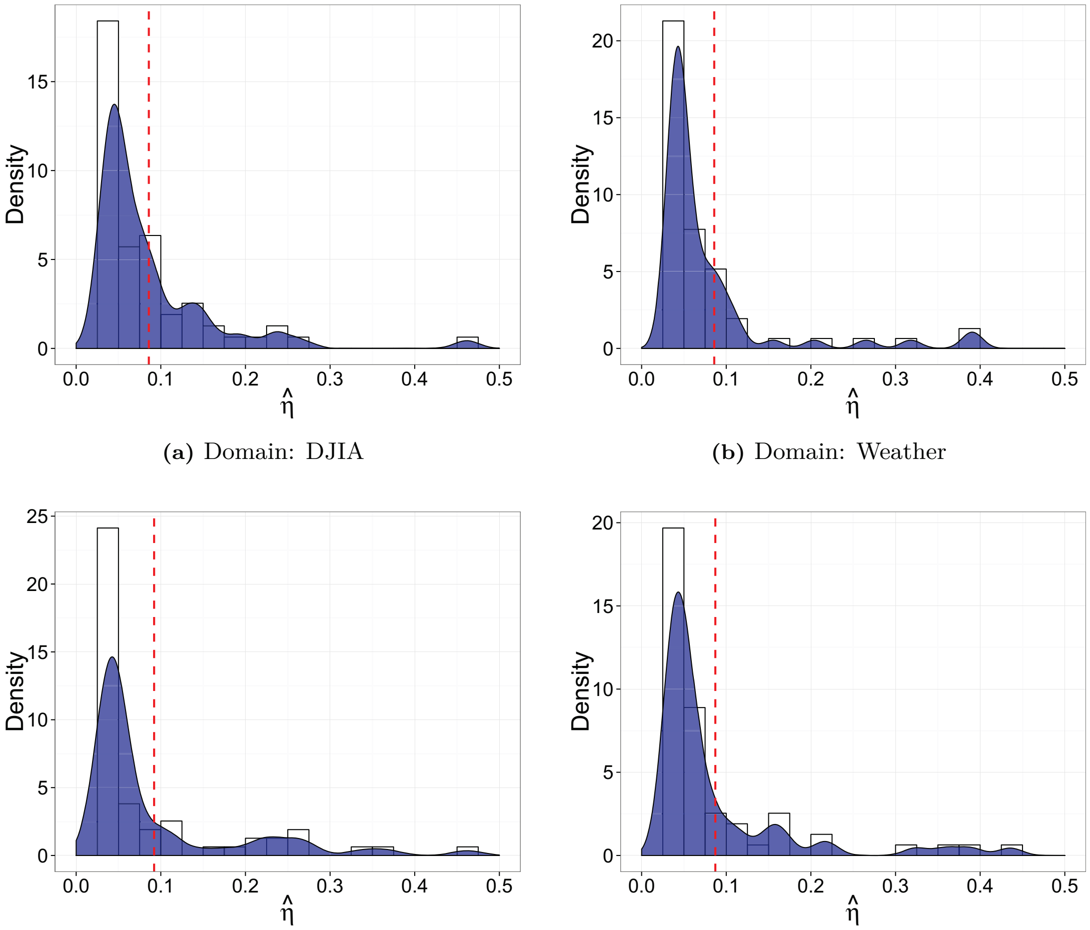
(a) Domain: DJIA (return in %)

(b) Domain: Weather (in °F)

(c) Domain: Gasoline Price (in US$/G)

(d) Domain: Diesel Price (in US$/G)

Figure A.10: Subject pool densities of MAP estimates $\hat{\sigma}$, with dashed lines illustrating the domain specific mean.



(a) Domain: DJIA

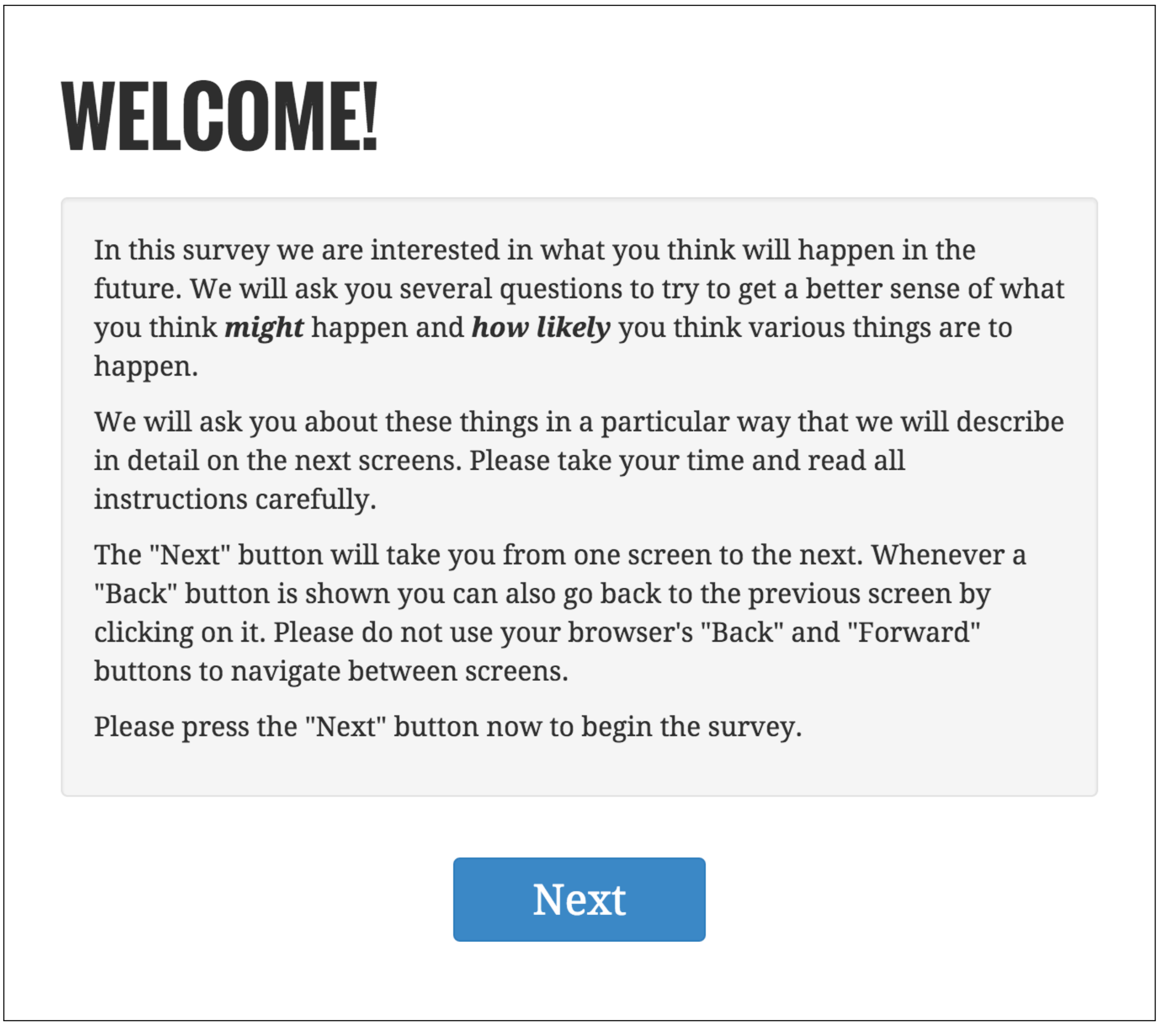
(b) Domain: Weather

(c) Domain: Gasoline Price

(d) Domain: Diesel Price

Figure A.11: Subject pool densities of MAP estimates $\hat{\eta}$, with dashed lines illustrating the domain specific mean.

A.3.5 Experimental Instructions (Screen-shots)



WELCOME!

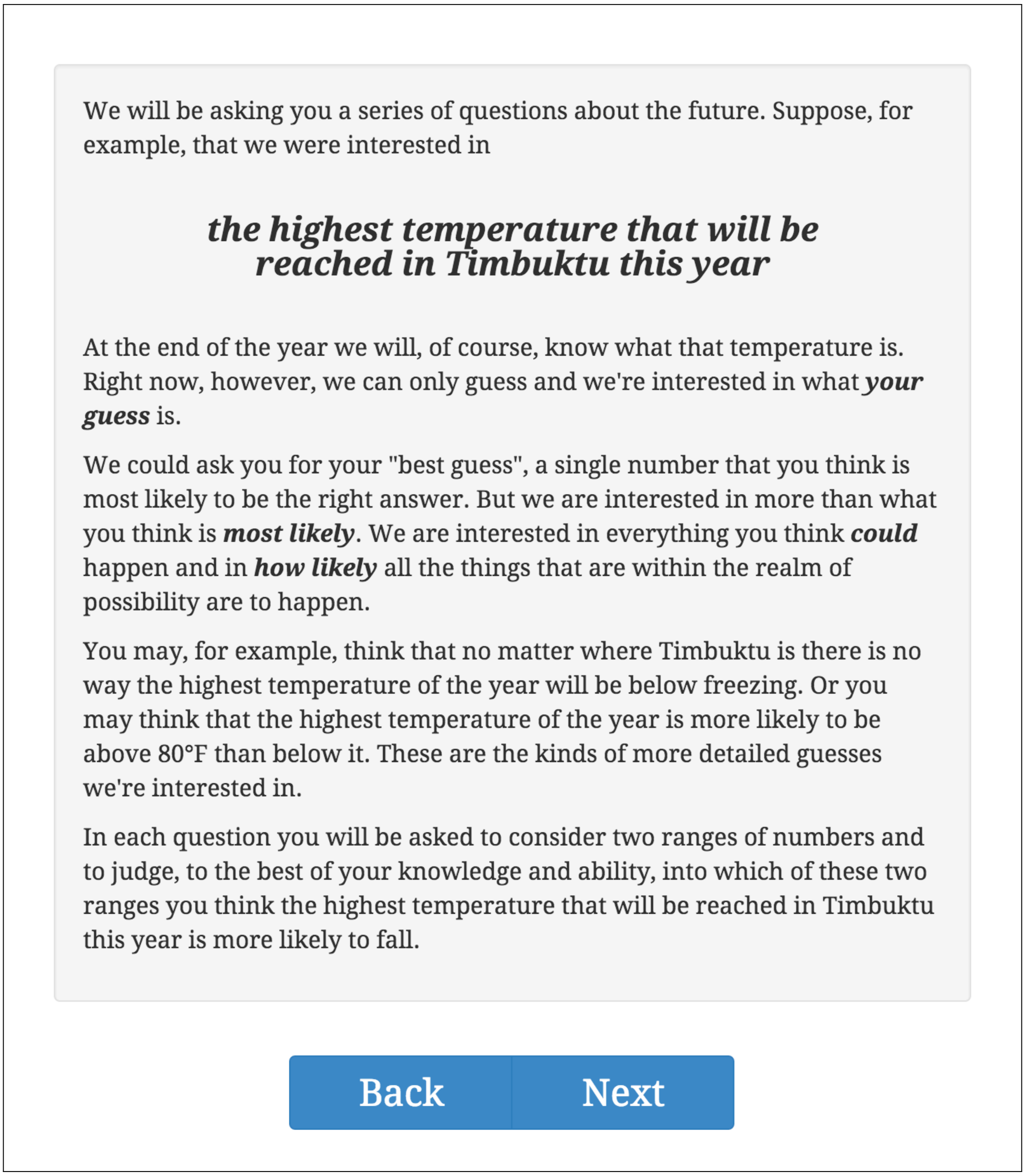
In this survey we are interested in what you think will happen in the future. We will ask you several questions to try to get a better sense of what you think ***might*** happen and ***how likely*** you think various things are to happen.

We will ask you about these things in a particular way that we will describe in detail on the next screens. Please take your time and read all instructions carefully.

The "Next" button will take you from one screen to the next. Whenever a "Back" button is shown you can also go back to the previous screen by clicking on it. Please do not use your browser's "Back" and "Forward" buttons to navigate between screens.

Please press the "Next" button now to begin the survey.

Figure A.12: Welcome screen.



We will be asking you a series of questions about the future. Suppose, for example, that we were interested in

the highest temperature that will be reached in Timbuktu this year

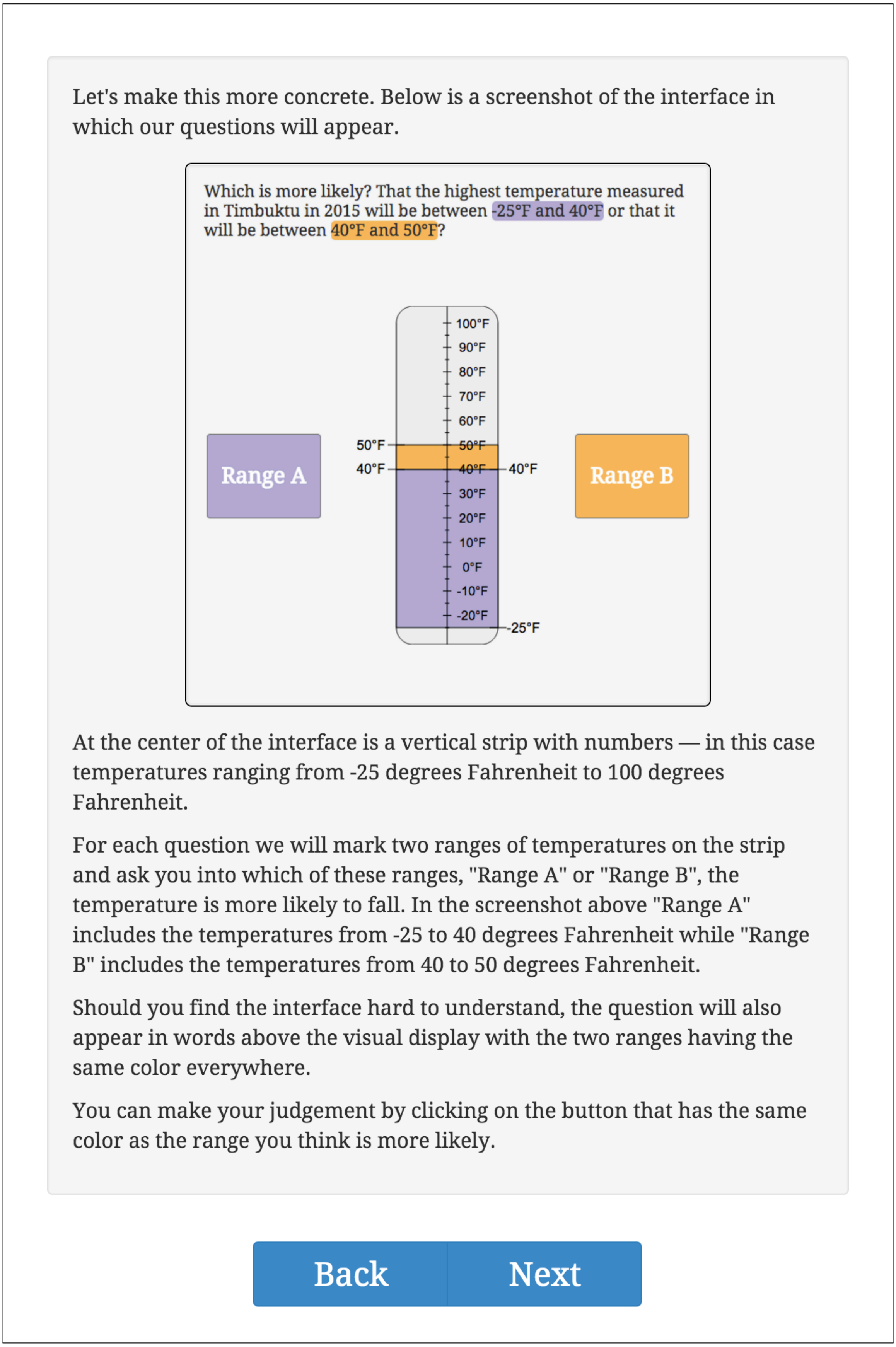
At the end of the year we will, of course, know what that temperature is. Right now, however, we can only guess and we're interested in what ***your guess*** is.

We could ask you for your "best guess", a single number that you think is most likely to be the right answer. But we are interested in more than what you think is ***most likely***. We are interested in everything you think ***could*** happen and in ***how likely*** all the things that are within the realm of possibility are to happen.

You may, for example, think that no matter where Timbuktu is there is no way the highest temperature of the year will be below freezing. Or you may think that the highest temperature of the year is more likely to be above 80°F than below it. These are the kinds of more detailed guesses we're interested in.

In each question you will be asked to consider two ranges of numbers and to judge, to the best of your knowledge and ability, into which of these two ranges you think the highest temperature that will be reached in Timbuktu this year is more likely to fall.

Figure A.13: Screen 1.



Let's make this more concrete. Below is a screenshot of the interface in which our questions will appear.

Which is more likely? That the highest temperature measured in Timbuktu in 2015 will be between -25°F and 40°F or that it will be between 40°F and 50°F?

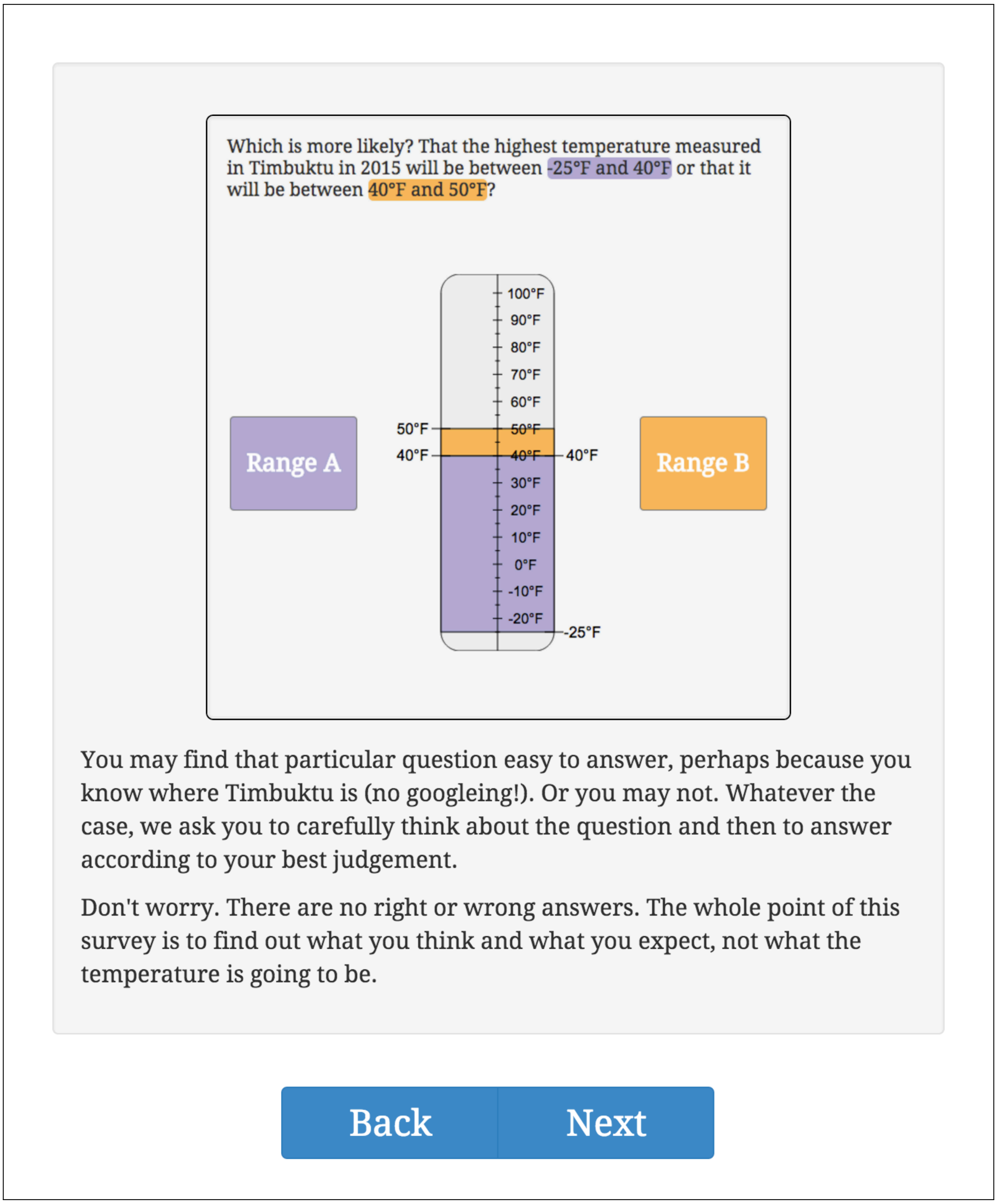
At the center of the interface is a vertical strip with numbers — in this case temperatures ranging from -25 degrees Fahrenheit to 100 degrees Fahrenheit.

For each question we will mark two ranges of temperatures on the strip and ask you into which of these ranges, "Range A" or "Range B", the temperature is more likely to fall. In the screenshot above "Range A" includes the temperatures from -25 to 40 degrees Fahrenheit while "Range B" includes the temperatures from 40 to 50 degrees Fahrenheit.

Should you find the interface hard to understand, the question will also appear in words above the visual display with the two ranges having the same color everywhere.

You can make your judgement by clicking on the button that has the same color as the range you think is more likely.

Figure A.14: Screen 2.

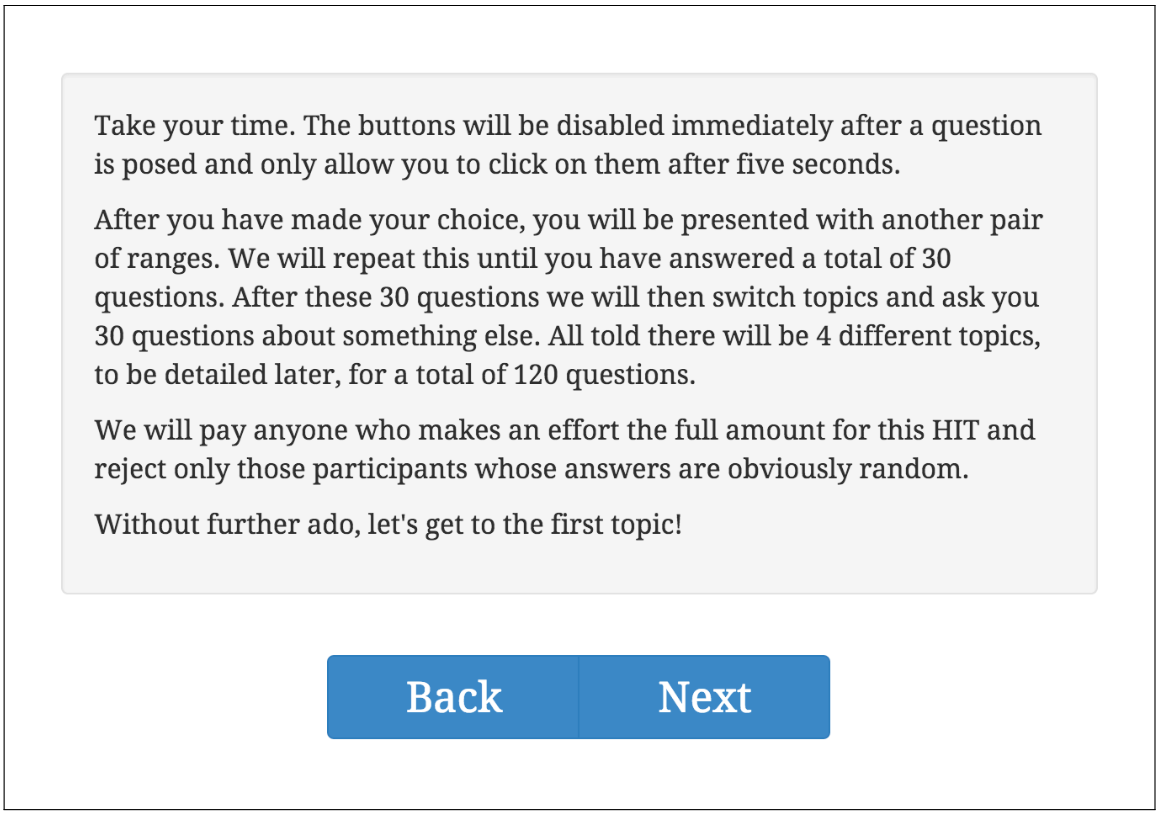


Which is more likely? That the highest temperature measured in Timbuktu in 2015 will be between -25°F and 40°F or that it will be between 40°F and 50°F?

You may find that particular question easy to answer, perhaps because you know where Timbuktu is (no googleing!). Or you may not. Whatever the case, we ask you to carefully think about the question and then to answer according to your best judgement.

Don't worry. There are no right or wrong answers. The whole point of this survey is to find out what you think and what you expect, not what the temperature is going to be.

Figure A.15: Screen 3.



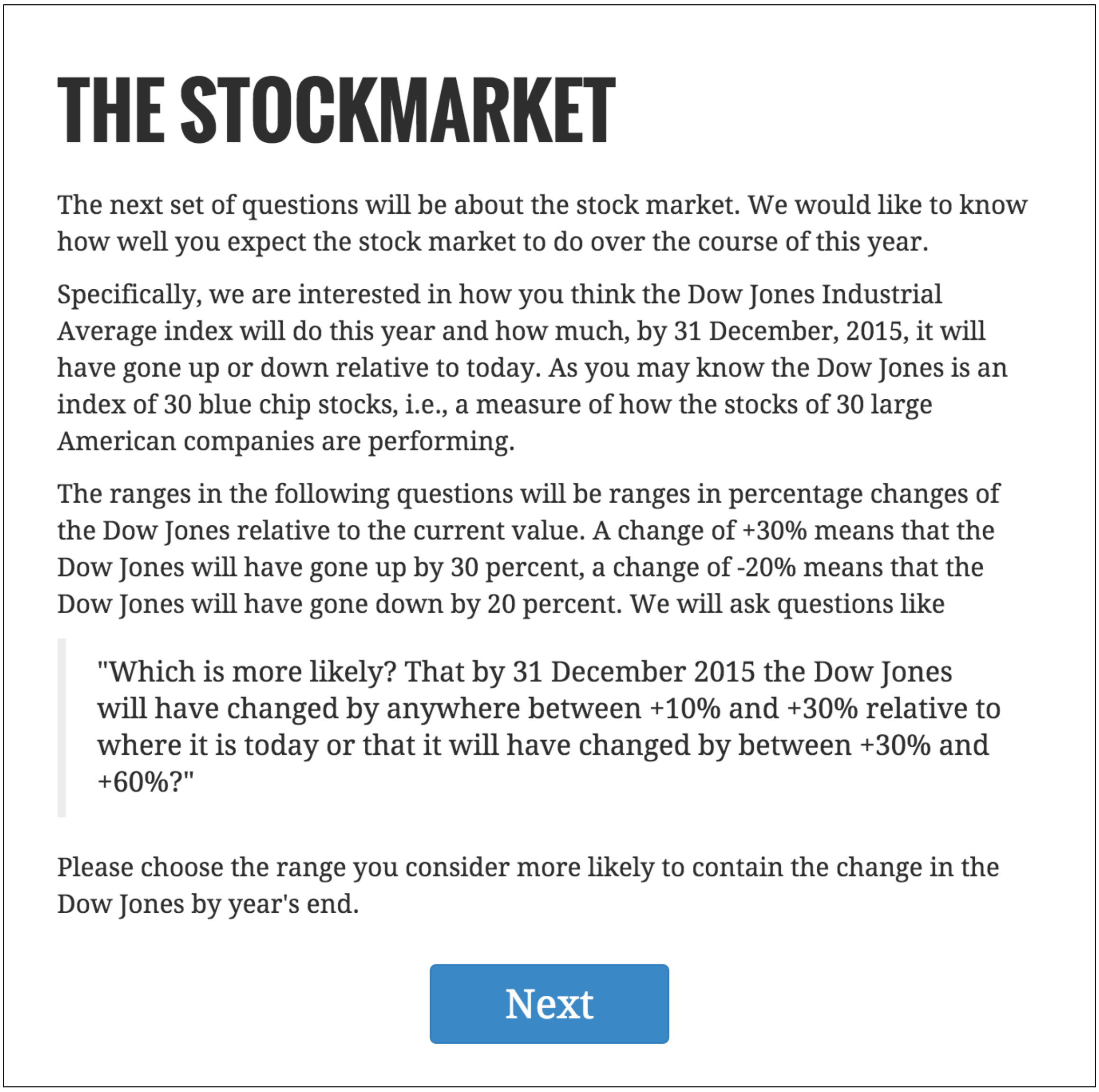
Take your time. The buttons will be disabled immediately after a question is posed and only allow you to click on them after five seconds.

After you have made your choice, you will be presented with another pair of ranges. We will repeat this until you have answered a total of 30 questions. After these 30 questions we will then switch topics and ask you 30 questions about something else. All told there will be 4 different topics, to be detailed later, for a total of 120 questions.

We will pay anyone who makes an effort the full amount for this HIT and reject only those participants whose answers are obviously random.

Without further ado, let's get to the first topic!

Figure A.16: Screen 4.



THE STOCKMARKET

The next set of questions will be about the stock market. We would like to know how well you expect the stock market to do over the course of this year.

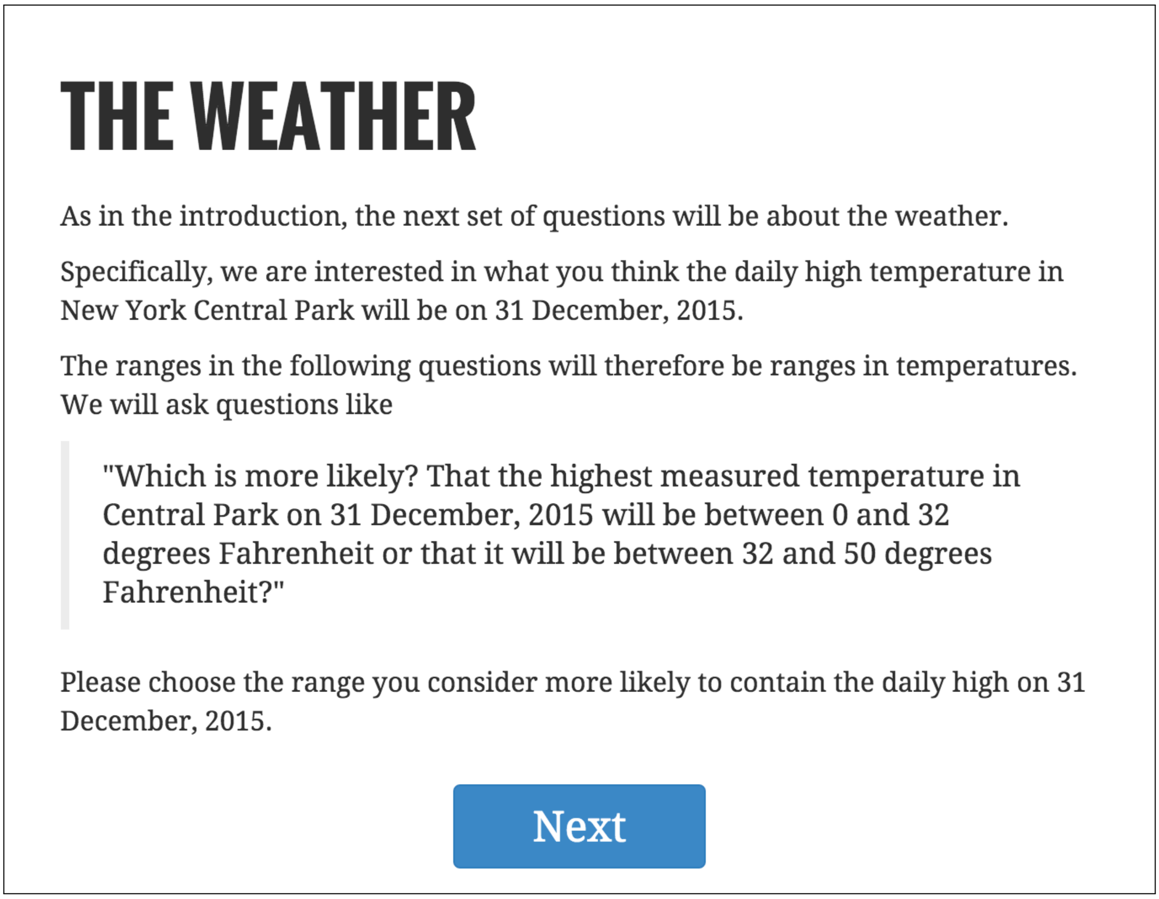
Specifically, we are interested in how you think the Dow Jones Industrial Average index will do this year and how much, by 31 December, 2015, it will have gone up or down relative to today. As you may know the Dow Jones is an index of 30 blue chip stocks, i.e., a measure of how the stocks of 30 large American companies are performing.

The ranges in the following questions will be ranges in percentage changes of the Dow Jones relative to the current value. A change of +30% means that the Dow Jones will have gone up by 30 percent, a change of -20% means that the Dow Jones will have gone down by 20 percent. We will ask questions like

> "Which is more likely? That by 31 December 2015 the Dow Jones will have changed by anywhere between +10% and +30% relative to where it is today or that it will have changed by between +30% and +60%?"

Please choose the range you consider more likely to contain the change in the Dow Jones by year's end.

Figure A.17: Domain explanation screen for DJIA domain.



THE WEATHER

As in the introduction, the next set of questions will be about the weather.

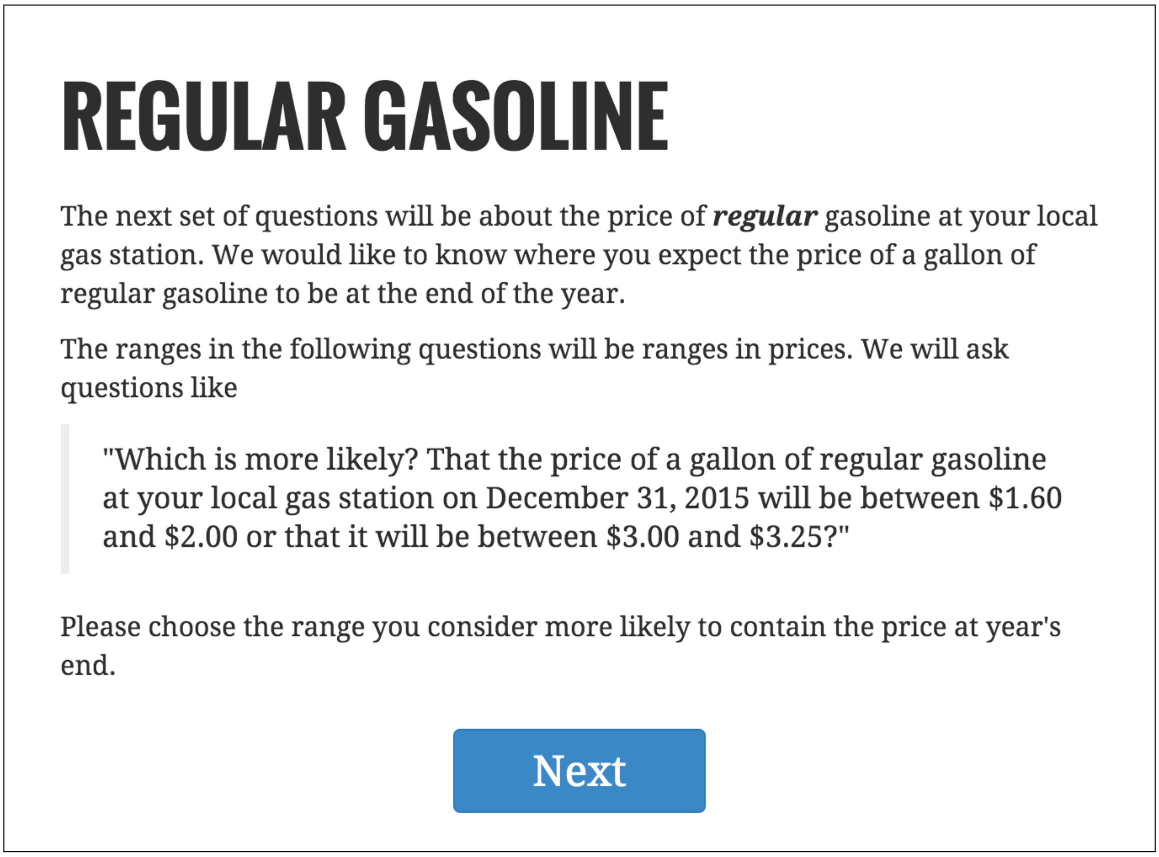
Specifically, we are interested in what you think the daily high temperature in New York Central Park will be on 31 December, 2015.

The ranges in the following questions will therefore be ranges in temperatures. We will ask questions like

> "Which is more likely? That the highest measured temperature in Central Park on 31 December, 2015 will be between 0 and 32 degrees Fahrenheit or that it will be between 32 and 50 degrees Fahrenheit?"

Please choose the range you consider more likely to contain the daily high on 31 December, 2015.

Figure A.18: Domain explanation screen for weather domain.



REGULAR GASOLINE

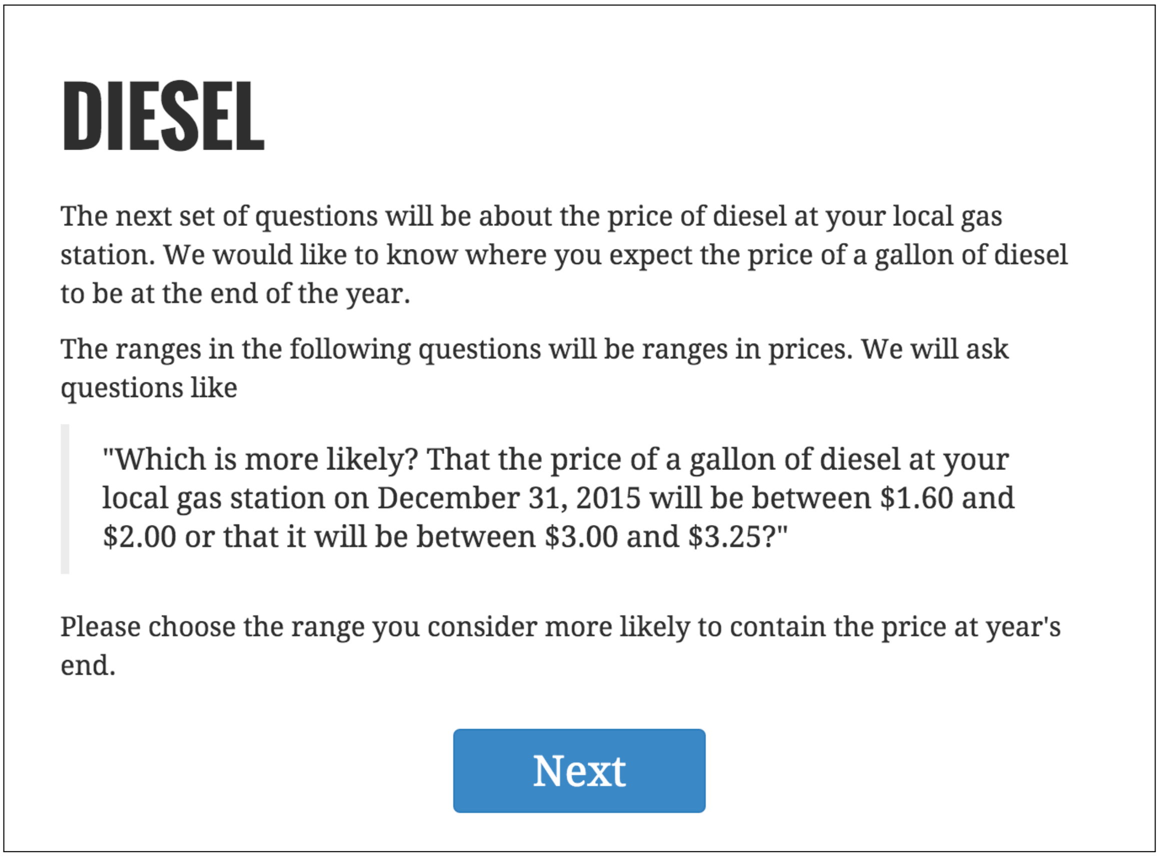
The next set of questions will be about the price of ***regular*** gasoline at your local gas station. We would like to know where you expect the price of a gallon of regular gasoline to be at the end of the year.

The ranges in the following questions will be ranges in prices. We will ask questions like

> "Which is more likely? That the price of a gallon of regular gasoline at your local gas station on December 31, 2015 will be between $1.60 and $2.00 or that it will be between $3.00 and $3.25?"

Please choose the range you consider more likely to contain the price at year's end.

Figure A.19: Domain explanation screen for gas domain.



DIESEL

The next set of questions will be about the price of diesel at your local gas station. We would like to know where you expect the price of a gallon of diesel to be at the end of the year.

The ranges in the following questions will be ranges in prices. We will ask questions like

> "Which is more likely? That the price of a gallon of diesel at your local gas station on December 31, 2015 will be between $1.60 and $2.00 or that it will be between $3.00 and $3.25?"

Please choose the range you consider more likely to contain the price at year's end.

Figure A.20: Domain explanation screen for diesel domain.

Ehrenwörtliche Erklärung

Das erste Kapitel der vorliegenden Arbeit basiert auf einem gemeinsamen Projekt mit Olga Nottmeyer, Ludwig Ensthaler und Georg Weizsäcker. Dieses Kapitel wurde außerdem von der Fachzeitschrift *Management Science* nach Abschluss eines Begutachtungsverfahrens zur Veröffentlichung akzeptiert. Das zweite Kapitel enstand aus einer Zusammenarbeit mit Georg Weizsäcker. Im Rahmen der Erarbeitung des dritten Kapitels habe ich in der frühen Phase des Projektes inhaltliche Unterstützung von Tobias Schmidt erhalten, Co-author der Endversion ist Herr Schmidt jedoch nicht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form in keiner anderen Prüfungsbehörde vorgelegt.

Ich bezeuge durch meine Unterschrift, dass meine Angaben über die bei der Abfassung meiner Dissertation benutzten Hilfsmittel, über die mir zuteil gewordene Hilfe sowie über frühere Begutachtungen meiner Dissertation in jeder Hinsicht der Wahrheit entsprechen.

Ort, Datum

Unterschrift