

# Three Essays in Economic Internet and Field Experiments

Inaugural-Dissertation  
zur Erlangung des Grades  
Doctor oeconomiae publicae (Dr. oec. publ.)  
an der Ludwig-Maximilians-Universität München

2012

vorgelegt von  
René Cyranek

Referent: Prof. Dr. Klaus M. Schmidt  
Korreferent: Prof. Dr. Martin G. Kocher  
Promotionsabschlussberatung: 15. Mai 2013

Datum der mündlichen Prüfung: 29. April 2013

Namen der Berichterstatter: Klaus M. Schmidt, Martin G. Kocher, Joachim K. Winter

*für Adrian*

# Danksagung

Ich bin vielen Personen zu Dank verpflichtet. Sowohl wissenschaftlich als auch persönlich waren sie wichtige Reisebegleiter.

Vorweg danke ich meinem Erstbetreuer Klaus Schmidt für seine Unterstützung. Ebenso meinem Zweitbetreuer Martin Kocher und meinem dritten Prüfer Joachim Winter. Diese Unterstützung brachte das Durchhaltevermögen, um die einer Promotion zugehörigen Talsohlen erfolgreich durchschreiten zu können. Es war eine bereichernde Erfahrung, über die Jahre hinweg Wissen zu akkumulieren und neue Erkenntnisse zu generieren.

Besonderer Dank gebührt meinen Koautoren Alain Cohn, Ernst Fehr, Ben Greiner, Michel Maréchal und Anthony Ziegelmeyer. Die räumliche Trennung hielt uns nicht davon ab, fruchtbar zusammenzuarbeiten.

In dieser besonderen Zeit als Doktorand lernte ich meine Mitstreiter in München sowie weitere Forscher aus aller Welt kennen. Im Gegensatz zum Vorurteil des sozial ungelinkten Wissenschaftlers waren diese Begegnungen überwältigend positiv. Zu mehreren Personen davon wird der Kontakt nie abreißen. Dafür bin ich schon jetzt dankbar.

Essentielle Hilfe bei der Organisation des Labors und bei der Durchführung von Experimenten habe ich durch viele tüchtige studentische Hilfskräfte erhalten. In den Jahren bei MELESSA haben wir erfolgreich den Betrieb aufrechterhalten und einander beschäftigt gehalten. Auch ihnen gegenüber möchte ich meinen Dank zum Ausdruck bringen. Des Weiteren danke ich allen anonymen Teilnehmern der Experimente.

Abschließend gebührt großer Dank meinen Eltern, auf die ich immer bauen konnte. Obwohl ich als Kind wie als Jugendlicher regelmäßig intensive Experimente an ihren Geduldsfäden durchführte, sind sie nie gerissen.

René Cyranek

# Contents

|   |          |
|---|----------|
| <b>Preface</b>  | <b>1</b> |
| <b>1 How to Pay in Internet Experiments?</b>  | <b>6</b> |
| 1.1 Introduction . . . . .  | 6        |
| 1.2 Related Literature . . . . .  | 9        |
| 1.2.1 Isolation, Reduction and Cross-Task-Contamination in the Random<br>Lottery Incentive Design . . . . . | 9        |
| 1.2.2 Stake size and other incentive effects . . . . .  | 12       |
| 1.2.3 Internet experiments vs. lab experiments . . . . .  | 14       |
| 1.3 Research Design . . . . .   | 15       |
| 1.3.1 General Setup . . . . .   | 15       |
| 1.3.2 Risk Elicitation . . . . .  | 19       |
| 1.3.3 Inconsistency . . . . .   | 22       |
| 1.3.4 Random Incentive Schemes & Ambiguity Aversion . . . . .   | 24       |
| 1.4 Hypotheses . . . . .  | 25       |
| 1.5 Results . . . . .   | 26       |
| 1.5.1 Descriptives . . . . .  | 26       |
| 1.5.2 Inconsistency . . . . .   | 29       |
| 1.5.3 Risk Aversion . . . . .   | 32       |
| 1.6 Conclusions . . . . .   | 42       |
| 1.7 Appendix . . . . .  | 44       |
| 1.7.1 Screenshots . . . . .   | 44       |
| 1.7.2 Ordered Probit Regression with Number of Safe Choices . . . . .                                       | 47       |

## CONTENTS

|          |  |           |
|----------|--|-----------|
| <b>2</b> | <b>What Makes them Tick</b>  | <b>50</b> |
| 2.1      | Introduction . . . . .   | 50        |
| 2.2      | Research Design . . . . .  | 55        |
| 2.2.1    | General Setup . . . . .  | 55        |
| 2.2.2    | Markets . . . . .  | 58        |
| 2.2.3    | Risk Elicitation . . . . .   | 60        |
| 2.2.4    | Beliefs . . . . .  | 61        |
| 2.3      | Hypotheses . . . . .   | 67        |
| 2.4      | Results . . . . .  | 68        |
| 2.4.1    | Descriptives . . . . .   | 68        |
| 2.4.2    | Data Imputation . . . . .  | 69        |
| 2.4.3    | Risk Aversion and Belief Correction . . . . .  | 70        |
| 2.4.4    | Comparison of Mean Beliefs, Equilibrium Prices, and Final Market<br>Prices . . . . . | 74        |
| 2.4.5    | Predictive Accuracy . . . . .  | 83        |
| 2.5      | Conclusions . . . . .  | 86        |
| 2.6      | Appendix . . . . .   | 89        |
| 2.6.1    | Screenshots . . . . .  | 89        |
| <b>3</b> | <b>Social Interaction</b>  | <b>97</b> |
| 3.1      | Introduction . . . . .   | 97        |
| 3.2      | Research Design . . . . .  | 100       |
| 3.3      | Theoretical Model . . . . .  | 102       |
| 3.3.1    | Mixed Room, No Information . . . . .   | 103       |
| 3.3.2    | Same Room, (No) Information . . . . .  | 104       |
| 3.3.3    | Mixed Room, Information . . . . .  | 106       |
| 3.3.4    | Discussion . . . . .   | 106       |
| 3.4      | Hypotheses . . . . .   | 108       |
| 3.5      | Results . . . . .  | 109       |

## CONTENTS

|       |                                |            |
|-------|--------------------------------|------------|
| 3.5.1 | Descriptives . . . . .         | 109        |
| 3.5.2 | Regressions . . . . .          | 110        |
| 3.5.3 | IAB-Data . . . . .             | 114        |
| 3.5.4 | Questionnaire Design . . . . . | 115        |
| 3.6   | Conclusions . . . . .          | 117        |
| 3.7   | Appendix . . . . .             | 118        |
| 3.7.1 | Input Mask . . . . .           | 118        |
| 3.7.2 | Payment Handout . . . . .      | 119        |
| 3.7.3 | IAB Questionnaire . . . . .    | 119        |
|       | <b>Bibliography</b>            | <b>121</b> |

# List of Tables

- 1.1 The ten paired lottery-choice decisions . . . . . 20
- 1.2 Demographic Characteristics of participants in the three experimental conditions . . . . . 27
- 1.3 Probit Estimates of Inconsistent Choice . . . . . 30
- 1.4 Results of Kolmogorov-Smirnov tests on differences in distribution of safe choices . . . . . 33
- 1.5 Ordered probit estimates (MSP) of likelihood to choose lottery A - models with full demographics . . . . . 35
- 1.6 Interval regression estimates of CRRA coefficient - models with full demographics . . . . . 39
- 1.7 Ordered probit estimates (Number of safe choices) of likelihood to choose lottery A - models with full demographics . . . . . 48
- 2.1 The ten paired lottery-choice decisions . . . . . 60
- 2.2 Demographic Characteristics of Participants . . . . . 68
- 2.3 Summary Statistics for Markets . . . . . 69
- 2.4 Estimates of  $r$  for different models . . . . . 71
- 2.5 Brier Scores of the 8 Matching Groups for Different Predictive Statistics and the set of all 16 Matches . . . . . 84
- 2.6 Brier Scores of the 8 Matching Groups for Different Predictive Statistics and the subset of 12 Matches without Germany . . . . . 85
- 3.1 Descriptives . . . . . 110
- 3.2 Regression of Output on Treatment Dummies and Ability . . . . . 112



## CONTENTS

|     |   |     |
|-----|---|-----|
| 3.3 | Regression of Output Difference on Treatment Dummies and Ability Difference . . . . . | 113 |
|-----|---|-----|

# List of Figures

- 1.1 Proportion of safe choices in each decision . . . . . 28
- 1.2 Predicted safe choices . . . . . 36
- 1.3 Screenshot of the Registration Page . . . . . 44
- 1.4 Screenshot of the Risk Aversion Elicitation Page . . . . . 45
- 1.5 Screenshot of the Risk Aversion Elicitation Page . . . . . 46
- 1.6 Predicted safe choices . . . . . 49
  
- 2.1 stated probability vs. subjective probability in a linear scoring rule for  
different degrees of constant relative risk aversion . . . . . 66
- 2.2 Stated Probability vs. Corrected Probability for Different Models of Risk  
Aversion Estimation . . . . . 73
- 2.3 Predictive Statistics vs. Mean Stated Belief for Different Models of Risk  
Aversion Estimation . . . . . 77
- 2.4 Predictive Statistics vs. Mean Corrected Belief for Different Models of  
Risk Aversion Estimation . . . . . 79
- 2.5 Screenshot of the General Rules Page . . . . . 89
- 2.6 Screenshot of the Belief Submission Page . . . . . 90
- 2.7 Screenshot of the Market Rules Page 1/3 . . . . . 91
- 2.8 Screenshot of the Market Rules Page 2/3 . . . . . 92
- 2.9 Screenshot of the Market Rules Page 3/3 . . . . . 93
- 2.10 Screenshot of a Market . . . . . 94
- 2.11 Screenshot of the Offer Submission Page . . . . . 95
- 2.12 Screenshot of the Bundle Trade Page . . . . . 96

## CONTENTS

|     |  |     |
|-----|--|-----|
| 3.1 | Ability Check . . . . .                              | 110 |
| 3.2 | Input Mask . . . . .                                 | 118 |
| 3.3 | IAB Questionnaire, English Version, page 1 . . . . . | 119 |
| 3.4 | IAB Questionnaire, English Version, page 2 . . . . . | 120 |

# Preface

This dissertation consists of three chapters that employ economic experiments as a method. All three chapters of this dissertation are self-contained and include their own introductions and appendices such that they can be read independently. The respective appendices contain the relevant supplementary material.

A famous quote from an economics textbook (Samuelson and Nordhaus, 1985) is “One possible way of figuring out economic laws ... is by controlled experiments. ... Economists (unfortunately) ... cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors. Like astronomers or meteorologists, they generally must be content largely to observe.”. This attitude has changed in the meantime. Experiments in economics are an established tool to answer questions which are otherwise hard to answer, because the researcher cannot obtain the relevant data or the data simply does not exist. The former, for example, is the case for trader characteristics in a financial market, while the latter is a frequent problem when economists consult politicians or companies. It is not always possible to deliver clear-cut predictions for new auction mechanism designs. So economic experiments serve as a wind tunnel to test new approaches before they are actually implemented.

Roth (1986) defines a general taxonomy to distinguish the reasons for economic experiments. The first category is “Speaking to Theorists”, which describes experiments conducted to test and refine economic theories. The second category is “Whispering into the Ears of Princes”, and contains experiments directly related to policy advice. The final category essentially catches all experiments that do not fit into one of the previous cat-

## PREFACE

egories. It is called “Searching for Facts” and has an exploratory focus. Its experiments try to gather data that deliver new insights into general economic questions. The three categories are not mutually exclusive. They are also not collectively exhaustive, although the bulk of economic experiments can be sorted into at least one of these categories.

The desirability of experiments comes from the fact that the researcher can control almost everything in such experiments. It may well be the case that the researcher does not regard all important factors in the development of a design or in the analysis of his results. But this can be changed for virtually all variables. As the experimenter controls the design and the participants of the experiment, he also has the potential to optimize all dimensions of the experiment, such that he can cleanly answer his research question. In this sense, economic experiments are very close to the experiments of chemists and other natural scientists.

While laboratory experiments were the main tool of analysis for most of the time, this has changed in recent years. Economists nowadays frequently conduct “field experiments” (Harrison and List, 2004). These are experiments, which change the established characteristics of laboratory experiments in order to obtain conclusions that are more robust than the conclusions from usual laboratory experiments. John List, a leading field experimenter from the University of Chicago, compares the usual lab setting to a clean test tube. But he is of the opinion that a dirty test tube can actually be preferable. “When the test tube is dirty, it means that it’s harder to make proper causal inference by using our typical empirical approaches that model mounds and mounds of data.”<sup>1</sup> This view is shared by many other researchers, so that field experiments have become an additional method to conduct economic experiments.

A step in between laboratory and field experiments are Internet experiments. Not all of them add field context, change the subject pool or use subjects unaware of the experiment. In some cases, Internet experiments are simply conducted in order to maximize

---

<sup>1</sup>Interview in a publication of the Federal Reserve Bank of Richmond  
[http://www.richmondfed.org/publications/research/region\\_focus/2012/q2-3/pdf/interview.pdf](http://www.richmondfed.org/publications/research/region_focus/2012/q2-3/pdf/interview.pdf)

## PREFACE

the number of participants. The first chapter of this thesis is a methodological analysis. It is based on joint work with Ben Greiner (UNSW Sydney) and Anthony Ziegelmeyer (Queen's University Belfast). It tries to answer the question, to which extent data generated in Internet experiments is similar to data generated in laboratory experiments. We implement an established payment scheme of the laboratory in the Internet and compare the participants' choices in lottery tasks to the choices of participants in auxiliary laboratory experiments. This payment scheme is called the random incentive system (RIS) (Baltussen et al., 2012) and its main feature is that only a subset of participants or decisions of a participant are actually paid out.

We discuss the results from this Internet experiment which aims to elicit the risk preferences of about 3,500 participants in the same way as Holt and Laury (2002) (HL) did in the laboratory. Only a subset of five of the Internet participants is paid for one out of ten choices. We compare the results to results from the laboratory treatments which implement a similar decision task but provide much higher probabilities of being paid. Specifically, in one laboratory treatment we use the same stakes as in the Internet, but a participant selection probability of  $1/15$ , while in another treatment we divide the stakes by 5, but pay  $1/3$  of participants according to one of their ten choices.

Our main result is that our Internet setting with an RIS elicits the same risk preferences through HL tables as a laboratory experiment with usual laboratory payment. This is the case although we implement an ambiguous probability of being selected for payoff. We find no differences between the risk preferences elicited in our main laboratory treatment and in the Internet treatment, which had higher stakes. However, we observe more inconsistencies in our Internet data than in our lab data.

The second chapter of this thesis is also an Internet experiment but does not ask a methodological question. It contributes to the literature on prediction markets. It is also based on joint work with Ben Greiner (UNSW Sydney) and Anthony Ziegelmeyer (Queen's University Belfast). Our markets were set up during the 2006 FIFA Soccer World Cup. The uncertain events to be predicted were the outcome of each of the matches.

## PREFACE

Each match had one independent market. We introduce the distinction between the (theoretical) equilibrium price and the (empirical) final price of a market. With our dataset of risk aversion, beliefs, and final prediction market prices we can compare the mean belief, the equilibrium price, and the final price on a market.

Our results support the prevalent notion of no relevant differences between mean belief and equilibrium price. Furthermore, our final prices exhibit a considerable difference to the equilibrium prices. Additionally, in our markets, the final prices are the worst predictive statistics and are outperformed by the mean belief. Our contribution to the literature is the distinction between final and equilibrium price as well as the comparison of the predictive power of mean belief, equilibrium price, and final price.

Finding out how prices on prediction markets are generated not only helps us to discover how prediction markets work. Through the similarity of prediction markets to regular double auction markets, analyzing these markets also aids our understanding of market price formation in general. Prediction markets have the advantage of richer datasets on traders and prices as usual financial markets. They also reveal the fundamental value of a contract at a certain point of time in contrast to many assets on regular financial markets.

Our experiment is the first study to analyze the interplay of risk aversion and beliefs with the aid of data elicited from human traders. We set up prediction markets which give us final prices to analyze. Additionally, we elicited the participants' risk aversion and beliefs. All other studies of the related strand of literature either relied on theoretical analyses or used (partly) artificially created datasets. Data created in actual prediction markets was never used in more but a supplementary way in this strand of literature. Furthermore, no study so far analyzes a dataset as exhaustive as ours.

The last chapter is a field experiment with the focus on a relative payment scheme and the channels through which such a scheme might yield employee collusion. It is based on joint work with Alain Cohn, Ernst Fehr, and Michel Maréchal (all three of them at the University of Zurich). Relative payment schemes are a useful tool to reward workers. The

## PREFACE

company knows its expenditures before the actual payment and is unaffected by positive production shocks. The worker in turn is unaffected by negative production shocks. But he is paid through a payment scheme, which is usually met with discomfort. A relative payment scheme is also susceptible to collusion. If workers find a coordination device, they could collectively lower their effort without changing their payment level.

We design a natural field experiment (Harrison and List, 2004) in order to find out whether social interaction or information drove the stated results of Bandiera et al.'s (2005) study. Our participants worked in dyads and were paid through a relative payment scheme. We vary the possibility of social interaction and the possibility of information exchange. Our results support the notion that social interaction reduces both output level and output difference of the workers. Performance information exchange in a dyad does not lead to a significant output reduction but to a reduction of the output difference. This leads to the conclusion, that the effect in Bandiera et al. (2005) was mainly driven by the social interaction channel.

This research adds to the literature of the effects of communication and social interaction on employees' output. In contrast to social interaction, which has a strong and clear reducing effect on both output level and output difference in a relative payment scheme, information on the coworker's output has a weaker and ambiguous effect on the output level and the output difference.



# Chapter 1

## How to Pay in Internet

## Experiments?

## Evidence from Risky Choices<sup>0</sup>

### 1.1 Introduction

The Internet provides a promising environment for experimental economic research. Its subject pool has a comparatively broad demographic distribution and consists of millions of individuals who only need to physically move towards a computer. Its lab space is never fully booked, its session schedule exactly matches the time preferences of the participants, and no session requires the presence of an experimenter. Nevertheless, the Internet did not supersede the laboratory as the premier environment for experiments without subject interaction. The most prominent concerns about Internet experiments revolve around data structurally different to lab data, incentive incompatible payment procedures, and more noisy data.

In this chapter, we address the question whether the Internet is a reliable experimental environment to employ a *random incentive system (RIS)* (Baltussen et al., 2012) for non-

---

<sup>0</sup>This chapter is based on joint work with Ben Greiner and Anthony Ziegelmeyer.

interactive, risky choices. We discuss the results from a large-scale Internet experiment which aims to elicit risk preferences of about 3,500 participants in the same way as Holt and Laury (2002) (HL) did in the laboratory. Only a subset of five of our participants is paid for one out of ten choices. We compare our results to results from two complementary laboratory treatments which implement a similar decision task but provide much higher probabilities of being paid. Specifically, in one laboratory treatment we use the same stakes and ten-choice list as in the Internet, but a participant selection probability of 1/15, while in another treatment we divide the stakes by 5, but pay 1/3 of participants according to one of their ten choices.

Many experiments pay a randomly selected part of participants' decisions, or even a randomly selected subset of participants. Such an RIS has been a widely used method in experimental economics since its beginnings.<sup>1</sup> Its advantage is that more choices can be elicited from participants. Specifically, it allows collecting decisions in different tasks from the very same subjects without inducing the problem of reference-point or wealth effects. These might occur if participants change their behavior in a later task due to income earned or lost in earlier tasks. Additionally, compared to paying all decisions by all participants, the random lottery design enables the researcher to increase the stakes in the single decision tasks while keeping the costs of the experiment at an affordable level.<sup>2</sup>

If a researcher conducts an experiment outside of the laboratory to increase the number of participants and without changing the payment scheme, the funds have to be increased accordingly. But the payment of thousands of participants is beyond the budget line of many researchers. Every such budget constrained researcher is interested in alternative payment schemes. This is exactly our research question. We want to find out, whether an

---

<sup>1</sup>In the literature, this kind of payment is often given different names, e.g. *random lottery incentive system* (Starmer and Sugden, 1991) or *random lottery selection method* (Holt, 1986). An extensive review and names list can be found in Baltussen et al. (2012).

<sup>2</sup>A related scheme is the *strategy method* (Selten, 1967), in which subjects submit full strategy profiles conditional on other players' choices. The profiles are then played out against each other, rather than making choices one by one during the course of play. This method allows to observe choice behavior at nodes which are rarely reached in pure play. The strategy method might suffer from similar problems as the lottery method: if a participant believes that a certain node will never be reached, his conditional decision in that node is not properly incentivized.

RIS to pay a subgroup of participants can be employed in the Internet if the goal is decision behavior similar to the one in the laboratory. We chose the domain of individual decision making because this domain is a natural field for Internet experiments, whose participants can participate according to their time preferences. The chosen tasks are decisions in risky choices because this field is well-researched and we can relate our research to it.

Internet experiments may suffer from the loss of control. Decision makers cannot be as closely monitored as in the lab, double participation and identity fraud are harder to tackle. Thus, we might expect that Internet experiments yield a larger amount of noise in the data, or even systematic biases. Two main critiques have been raised regarding the random lottery design. Holt (1986) argues that if (1) rather than isolating the different decision tasks, participants reduce the whole experiment to a single, complex decision problem, and (2) participants' decisions do not follow the independence axiom of expected utility theory (EUT), then the lottery design can generate spurious results in the sense of cross-task contamination of decisions. Harrison (1994) points out that in a random lottery design, the announced incentives for each task are diluted by the fact that each task has only a small probability of being for real. Not properly incentivizing decisions might yield more noise and biases in the collected data.

Both critiques have inspired a number of papers testing for cross-task contamination in the random lottery method and incentive effects in experiments (see our detailed discussion in Sections 1.2.1 and 1.2.2, respectively). The prevalent result seems to be that Holt's (1986) concern is unsubstantiated: while indeed many people violate the independence axiom (see the famous Allais paradox of Allais (1953)), there is no evidence that they reduce single tasks in the random lottery design to one complex decision problem. Increasing stake sizes usually change the results of individual decision tasks and social interactions quantitatively, but does not suppress the qualitative behavioral pattern.

Our main result is that our Internet setting with an RIS elicits the same risk preferences through HL tables as a laboratory experiment with usual laboratory payment. This is the case although we implement an ambiguous probability of being selected for payoff.

We find no differences between the risk preferences elicited in our main laboratory treatment and in the Internet treatment, which had higher stakes. However, we observe more inconsistencies in our Internet data than in our lab data.

## 1.2 Related Literature

In this section we review the literature on cross-task contamination in the random incentive system (1.2.1), stake size effects in experiments (1.2.2), and general effects of using the Internet rather than the laboratory (1.2.3). For this purpose, it is helpful to follow the terminology used in Baltussen et al. (2012) and to distinguish between within-subjects random lottery designs (WS, only one randomly selected out of several tasks is paid), between-subjects random lottery designs (BS, only a randomly selected subset of participants is paid), and hybrid designs (WS & BS). Note that the cross-task contamination critique only applies to designs with a WS dimension, while incentive dilution might be an issue in all random lottery designs. In general, we will focus on individual decision experiments, and will discuss interactive decision situations only marginally.

### 1.2.1 Isolation, Reduction and Cross-Task-Contamination in the Random Lottery Incentive Design

In a random lottery incentive design, each experimental participant is confronted with a number of tasks. Each task can be a choice between strategies, risky prospects, outcome distributions, etc. Depending on choices of the participant and choices of others, each task yields a certain monetary outcome (distribution). However, as publicly known to all participants before the experiment, only one of the tasks will be randomly selected for actual payment. Let us define “true” preferences with respect to a specific task as those preferences that would be elicited if the participant would face this certain task only. Then the question is whether a participant in a random lottery incentive design reveals

the same ‘true’ preferences as he would in the single task.

That people violate the independence axiom in their decisions is no news for experimental economists and psychologists. The plenty evidence for the Allais paradox is just one example. The question remains whether participants indeed reduce the random lottery tasks to a single choice, and this can be tested experimentally.

### **Within-Subjects Design (WS)**

Starmer and Sugden (1991) provide experimental evidence according to which the reduction hypothesis in the strong form can be rejected. But they do not establish that the random lottery design is immune to cross-task contamination.<sup>3</sup> In order to be sure about contamination, they propose further studies with a larger number of subjects. This has been done by Beattie and Loomes (1997) and Cubitt et al. (1998), who do not find contamination. The hypothesis that there is no difference between responses to random lottery and single choice experiments seems adequate to organize the data in the case of binary choice among simple lotteries.

Confirming these results, Laury (2005) does not find different behavior when comparing subjects who were paid according to only one row of an HL table and subjects who were paid according to all rows. With this finding in mind, we see our approach of paying only one row of the HL table for a subject justified.

Hey and Lee (2005a) analyze whether subjects in an RIS context separate questions and Hey and Lee (2005b) test whether subjects are influenced in their decisions by their past behavior. Both results provide reassuring support for experimental economists. Subjects seem to separate and are not influenced by past decisions.

Similarly, Camerer (1989) implements several decisions of lottery pairs for each subject. After a lottery pair was determined for payment, he asked subjects whether they wanted to revise their stated preference in the two lotteries. Only 2 of 80 subjects changed their

---

<sup>3</sup>Indeed, the reduction hypothesis represents the extreme case of such contamination, just as the isolation hypothesis represents the opposite extreme case in which there is no contamination at all.

preference. This is either support for the isolation effect or the independence axiom. As his data violates the independence axiom, he concludes that these results support the isolation effect.

In a paper that is mainly concerned with cognitive costs of decision making Wilcox (1993) finds that the probability of task selection is not important if choices concern simple lotteries. But still, a higher probability of task selection changed subjects' behavior in a complex setting. We do not see HL tables as complex (and also claim a successful job of our instructions) and support this perspective by the very low share of inconsistencies in our data.<sup>4</sup>

Overall, most results speak in favor of no contamination or reduction. However, Baltussen et al. (2012) find support for contamination from one task to the following one. As the task analyzed in this chapter is the first one in the whole experiment, we do not have to regard such contamination.<sup>5</sup>

### **Between-Subjects Design (BS)**

A BS design in the risk domain has been employed by Harrison et al. (2007b) and Harrison et al. (2007a). Both studies deal with risk aversion in different field contexts. They employ HL tables as a workhorse. The former paper estimates the risk aversion of a representative sample of Danes, each of whom is only paid with a 10% probability. The latter paper compares HL-elicited risk aversion to risk aversion elicited in a field context. HL tables serve as a good proxy for risk aversion as long as the risk aversion in the field does not have background risk.

Camerer and Ho (1994) analyze risk behavior in a laboratory setting. Their main focus is on the betweenness axiom and nonlinear probability. In a sub-part of their experiment,

---

<sup>4</sup>We find inconsistencies for 5% of the lab subjects and 14.3% for the Internet subjects. Compared to the lab studies of Holt and Laury (2002) with 19.8%, Maier and R uger (2011) with 24%, Bruner et al. (2008) with 30%, or Jacobson and Petrie (2009) as well as Prasad and Salmon (2010) with more than 50% of inconsistent responses, only very few of our subjects behaved inconsistently.

<sup>5</sup>This very first task consists of ten subsequent decisions. So there might be carry-over effects within that one task. Unlike Baltussen et al. (2012) we do not provide participants with feedback in between these ten decisions. This way, the carry-over effects are minimized. Additionally our main tool of analysis – ordered probit regression – regards potential interdependencies of the ten decisions.

they play out gambles for one of 36 subjects and find sound results of the elicited risk aversion. Also in the lab, Baltussen et al. (2012) find a lower risk aversion in the BS treatment and offer the computational ease of lottery reduction in this treatment as an explanation.

There is general support for employing a BS design in an RIS. Although, there seem to be some unresolved issues about the exact context in which BS yields unbiased results. Further research should be stimulated.

### **Hybrid Designs (WS & BS)**

In the paper mentioned in the BS paragraph, Harrison et al. (2007b) analyze risk attitudes in Denmark. They run extra treatments in order to control for this issue and do not find differences between a WS design and a hybrid design (footnote 16). In another large scale experiment Harrison et al. (2002) elicit discount rates from a representative pool of Danes and do not find differences in discount rates between the treatments with an implementation probability of 0.25 and an implementation probability of 1.

Another study in an ultimatum bargaining context – Armantier (2006) – compares the data of WS and BS designs and does not find differences. Stahl and Haruvy (2006) analyze the RIS in a dictator game setting and find significant differences when it is employed. This is achieved through path dependency, which gives potential for warm-glow effects.

Like pure WS or pure BS settings, a hybrid design does not distort the participants decisions in risk aversion or discount rates domains. There seem to be interaction effects in studies with a social dimension. As we can rule out direct other-regarding preferences in our setting, we consider the latter results as negligible in our context.

### **1.2.2 Stake size and other incentive effects**

Harrison (1994) points out that in a random lottery experiment, the apparent incentives offered by the face values of the options are diluted by the fact that each task has only a

small probability of being for real. Thus, the *expected* money payout per subject per task is usually very small. Harrison argues that in such cases, the random lottery design is biased towards those responses that are most likely to result through erroneous behavior: for a given task, random lottery responses will contain more errors than single choice responses. If we are to test this claim, we need to combine the principal hypothesis, that the frequency of errors is negatively related to the strength of incentives, with some auxiliary hypothesis about the properties of true preferences and/or the nature of errors. One such auxiliary hypothesis is that true preferences satisfy the axioms of EUT, and that the violations of EUT found in experiments are a product of the weak incentives for correct reasoning, offered by those experiments. On this view, systematic deviations from EUT result from subjects' using simplifying heuristics to economize on mental effort. The stronger the incentives associated with a decision task, the less such heuristics will be used. One implication of this hypothesis is that violations of EUT should be more pronounced in random lottery experiments than in single choice experiments. Wilcox (1993) presents a version of this argument. He hypothesizes that the greater the dilution of incentives in a random lottery design, the less "accurate" will be the heuristics used by subjects. He reports an experiment whose results provide support for his hypothesis when decision tasks require choices over compound lotteries, but no support in the case of choices over simple lotteries. In other words, task complexity and strength of incentives interact.

Another auxiliary hypothesis is that subjects' risk aversion is positively related to the strength of incentives. In a lottery choice experiment, Holt and Laury (2002) report that scaling up payoffs (by factors of 20, 50, and 90) causes a significant increase in risk aversion (see also Harrison et al., 2005 and Holt and Laury, 2005).<sup>6</sup> Therefore, if paying all choices (instead of one chosen at random) is interpreted by subjects as an increase in

---

<sup>6</sup>Note, however, that while Holt and Laury (2002) find a stake effect between  $1\times$  and higher stakes, there are no significant differences in risk preferences among the three high stake conditions in their data. Potential reasons include missing statistical power or some kind of adjustment to the "real" level of risk aversion which cannot be increased by higher stakes any more.



the expected payoff for each decision, one would expect to see more risk aversion among subjects who are paid for all choices instead of one choice selected at random.

Laury (2005) reports a lottery choice experiment, where subjects are presented the same choice tasks as in Holt and Laury (2002). She tests whether subjects behave as if each of these choices involves the stated payoffs, or if subjects scale-down payoffs to account for the random selection that is made. Three treatments are conducted: pay for 1 of 10 choices under low payoffs, pay for 10 of 10 choices under low payoffs, and pay for 1 of 10 choices under 10x the low payoff level. Increasing payoff scale has a significant effect on choices compared with the low payoff treatments where all 10 decisions are paid, or where one decision is paid. However, there is no significant difference in choices between paying for 1 or all 10 decisions at the low payoff level. This again supports the validity of the random lottery incentive system in the case of choice among simple lotteries.

Summarizing the results from the given studies high payoffs “sharpen” the results of experimental research compared to the usual lab payoffs. They at least reduce variance around an (expected) theoretical outcome and often enough shift the results to this outcome. In the context of risk aversion, they make this aversion more pronounced. High stakes do not rule out irrational behavior.

### **1.2.3 Internet experiments vs. lab experiments**

When leaving the lab as an environment for the experiment, the experimenter loses some control, as she cannot control the participants’ behavior as well as in the lab any more. For example multiple participation in the laboratory can be prevented to a much higher extent than in the Internet, where individuals have a positive chance to pass various checks and register more than once in the internet section of our experiment. On the positive side are the increased number of participants and the broader variety of their characteristics.

Anderhub et al. (2001) provide – to the best of our knowledge – the first comparison between internet and laboratory experiments in economic decision making. They run an experiment concerned with individual decision making in the field of inter-temporal choices. Internet and lab environments generated similar data with Internet data showing the usually higher variance.

The Internet compared to the laboratory as an experimental environment seems to have an effect similar to low stakes compared to high stakes. It increases variance of the data, moves it (further) away from theoretical predictions, and reduces risk aversion. However, it is important to note that in general terms the data is still fine. Internet participants submit similar decisions as lab participants.

To fully exploit the possibilities of Internet experiments the researcher has to implement an RIS. Only this provides her with the possibility of recruiting more participants than in the lab without paying for each and every single one of them. Potential obstacles are interaction effects between the RIS and the Internet environment. So before an experimental setting with an RIS in the Internet is declared as a sound method, it has to be tested. Does this setting provide the researcher with the same results a regular lab setting would? To the best of our knowledge there is no research on decision making in such a setting. We provide results from the domain of risky choices. Our results support the notion that the Internet and RIS are on good terms with each other. At least for the domain of risk aversion, RIS and the Internet can be used together.

## **1.3 Research Design**

### **1.3.1 General Setup**

The risk preferences elicitation experiment we describe in this chapter was embedded in a larger research project on the processing of information in parimutuel and double auction prediction markets. As the underlying event for the experiment we chose the FIFA

Soccer World Cup 2006. Our Internet session with 3,582 participants is an “artefactual field experiment” in the terminology of Harrison and List (2004): it has a nonstandard subject pool but still employs an abstract framing and an imposed set of rules.<sup>7</sup> In order to compare this dataset to lab data, we ran additional laboratory sessions. 120 individuals in 8 sessions took part in the Cologne Laboratory for Economic Research (CLER). Participants were recruited from those in the CLER subject pool who did not participate in the Internet experiment. Sessions lasted approximately one hour. Additional to their earnings from the experiment, laboratory participants received a showup fee of 2.50 €. All participants selected for payment were informed by e-mail and paid by wire transfer after the World Cup final took place.

Participants of the Internet experiments were recruited by inviting individuals from several subject pools of experimental laboratories in Germany.<sup>8</sup> This way we did not fully exploit one of the advantages of Internet experiments - diversity in subject characteristics. But this gave us similar subject pools across our Internet and lab experiments, which will later on help in our analysis.<sup>9</sup> In order to demonstrate the serious scientific background as well as the financial capacity to pay the participants’ winnings, the experiment homepage prominently informed that the experiment was conducted by researchers of the University of Cologne and the Max Planck Institute of Economics in Jena.

The entire webpage was presented in German and in English. In a first step of the registration process, participants filled in a form which asked for name, e-mail address, a chosen username and password, and then received an e-mail with a link to verify and

---

<sup>7</sup>For the extent of “artefactuality” of the online session see the next paragraph and the descriptives in Section 2.4.1.

<sup>8</sup>We gratefully acknowledge the support of the experimental laboratories at the University of Bonn, the University of Cologne, the University of Erfurt, Humboldt-University of Berlin, the Technical University of Berlin, the Max-Planck-Institute of Economics in Jena, the University of Magdeburg, and the University of Mannheim. All laboratories used the recruitment system ORSEE (Greiner, 2004) for approaching participants by e-mail. Additionally we sent e-mails to mailing lists at the University of Cologne and posted links at the university’s web pages. E-mail recipients could forward the invitation without invalidating the registration link. Participants could also register by directly accessing the experiment homepage [www.torlabor.de](http://www.torlabor.de).

<sup>9</sup>82% of our participants stated that they reacted to an invitation e-mail, while 16% registered due to the recommendation of a friend, and the remaining 2% came through other channels.

complete their registration.<sup>10</sup> The next step was to make choices in the HL table described below. This was followed by a questionnaire for demographical data like year of birth, job, student and employment status, field of studies, etc.<sup>11</sup> Screenshots of the registration page and the demographic questionnaire can be found in the appendix. After this registration procedure, participants could enter the belief elicitation tasks and the markets for the 64 world cup games.

In both the Internet experiment and the corresponding laboratory sessions, taking part in this risk preferences elicitation procedure was the first of three tasks participants had to complete. In all conditions, the second task was to state probabilistic beliefs about possible outcomes of the soccer matches in the FIFA World Cup 2006, and the third task consisted of trading contracts that yield payments based on the outcomes of these soccer matches. In the Internet treatment, the second and third task were completed over the course of the World Cup, for 64 matches played in four weeks. The lab sessions lasted for approximately one hour and the subject for the second and the third task was the final between Italy and France. In this chapter, we focus on the first task, which is the risk preferences elicitation described below. To keep the experiment environment as similar as possible, all experimental instructions were identical in the Internet and in the laboratory. In the Internet, risk choices were made in a web form, while they were collected by pen and paper in the lab.

As already described, we set up two main treatments - Internet and laboratory. The Internet treatment deals with relatively high stakes and a low probability of being picked for payment. The laboratory treatment has usual lab stakes and a much higher probability

---

<sup>10</sup>The primary key to differentiate participants was their e-mail address. It was impossible to register multiple times with the same e-mail address. Still participants could circumvent this check by using different e-mail addresses. To minimize multiple registrations, it was explicitly noted that this would lead to immediate exclusion from the experiment and all payments. Regularly we conducted spot tests. As none of our checks yielded a cheating participant we never had to exclude a participant because of multiple registrations.

<sup>11</sup>Most answers were voluntary. We emphasized that answering the voluntary questions and answering truthfully would strongly support our work as researchers. But we also encouraged participants to skip questions they were uncomfortable with, so they would not quit their registration out of privacy reasons. We did not elicit the full set of demographics from all participants, as maximizing the number of participants was our main goal.

of being chosen for payment.<sup>12</sup> These two treatments resemble the usual characteristics of laboratory and Internet experiments. As the Internet session was part of a larger project, we could not vary its dimensions of payment probability or stake size. And introducing a laboratory treatment which mimics the payment probability of the Internet treatment (0.14%) would have been infeasible. Therefore, as an auxiliary treatment we introduce a laboratory treatment in which we reduce the payment probability but increase the stake size to the Internet stake size. This leaves us with three different treatments overall: lab regular stakes (60 participants), Internet (3,582 participants), and lab high stakes (60 participants).

To motivate participants we used an RIS. In the Internet, five of all participants were paid according to one of their choices in the risk preference elicitation task.<sup>13</sup> However, there was no hint at all on the web pages about the total number of participants. As a downside, we generate ambiguity about the payment probability. This adds a new dimension to the analysis. But as an upside, our expected payment was orthogonal to the number of participants as we only paid a certain number of them regardless of the overall participant number. The ex post probability to be selected for payment in the risk preference elicitation task was  $5/3582=0.14\%$ . As only one of ten choices of the MPL was paid, the probability that an individual choice mattered was 0.014%. Due to the nondisclosure of the participant number, no subject knew about this figure.

An alternative to this approach would have been to pay a fixed number per 100 participants. This procedure would not have let ambiguity influence our results. But this way, the expected payoffs would have been related to the number of participants. As this Internet experiment was the first one using this recruiting procedure we had no sound estimate of the potential participant number.

In the lab high stakes payoff scheme, the associated payoffs of lottery choices were the

---

<sup>12</sup>Actually, the probability is 1 as everybody in this treatment was paid. Not all participants were paid for the same task, though.

<sup>13</sup>Another 5 were paid according to their belief statements, and another 20 according to their market performance. It was emphasized in the instructions of each task that participants who were paid according to one task would not be paid according to another one.

same as used in the Internet and shown in Table 1.1. A screenshot of the table displayed to the participants can be found in the appendix. Only one of the 15 participants in each session was paid according to one of her HL table choices. Thus, moving the experiment from the Internet into the lab increased the objective probability that an individual HL table choice mattered to  $1/15 * 1/10 = 0.667\%$ , which is almost 50 times higher than in the Internet. More important than that, the participants in both lab payoff schemes had no ambiguity about the payment probability. Under the lab regular stakes payoff scheme, we changed the monetary lottery payoffs and the selection probability, but kept the expected money payout per decision constant.<sup>14</sup> Specifically, 5 randomly selected out of the 15 participants were paid according to one of their HL table choices, but all monetary lottery payoffs in Table 1.1 were divided by 5. Thus, the probability that one of the lab regular stakes HL decisions mattered was now  $1/3 * 1/10 = 3.333\%$ .<sup>15</sup>

### 1.3.2 Risk Elicitation

Our workhorse in this chapter is a simple experimental measure for risk aversion called *multiple price list* (MPL), introduced by Holt and Laury (2002). This approach goes back to a method originally presented in section 6.2(a) of Farquhar (1984). The HL table has been heavily used in laboratory experiments. The method involves a relatively transparent task: Each subject is presented with a choice between two lotteries which we call *A* and *B*.<sup>16</sup> Table 1.1 illustrates the basic payoff matrix presented to subjects in our Internet experiment.<sup>17</sup> The first row shows that lottery *A* offers a 10% chance of receiving 100€ and a 90% chance of receiving 80€. Similarly, lottery *B* in the first row has the same probabilities, but for payoffs of 192.50€ and 5€, respectively. Thus the two lotteries

---

<sup>14</sup>In the HL tables we kept the numerical amounts associated with the lotteries identical. Instead of Euro we used ECU (Experimental Currency Units) in the lab, with publicly known conversion rates of 1€ (0.2€) for 1 ECU, respectively, in the high (regular) stakes condition.

<sup>15</sup>In the lab high stakes payment scheme, two further participants in each session were selected for payment of the second and third task, respectively. Under the lab regular stakes scheme, 5 of the 10 remaining participants were paid for the second task, and the last five for the third task. The lab regular payoff scheme represents the usual application of a within-subjects RIS.

<sup>16</sup>In their analysis, HL coined the descriptions “safe” and “risky” for lotteries *A* and *B*, respectively.

<sup>17</sup>At the time of the experiment, the exchange rate of 1€ was about US\$1.25.

INTERNET PAYMENT

Table 1.1: The ten paired lottery-choice decisions

| Row | Lottery $A$              | Lottery $B$                | $E_A - E_B$ | CRRA coefficient if row was last A choice, below all B choices |
|-----|--------------------------|----------------------------|-------------|--|
| 1   | { 100€, 0.1 ; 80€, 0.9 } | { 192.50€, 0.1 ; 5€, 0.9 } | 58.25€      | [-1.71; -0.95]   |
| 2   | { 100€, 0.2 ; 80€, 0.8 } | { 192.50€, 0.2 ; 5€, 0.8 } | 41.50€      | [-0.95; -0.49]   |
| 3   | { 100€, 0.3 ; 80€, 0.7 } | { 192.50€, 0.3 ; 5€, 0.7 } | 24.75€      | [-0.49; -0.14]   |
| 4   | { 100€, 0.4 ; 80€, 0.6 } | { 192.50€, 0.4 ; 5€, 0.6 } | 8.00€       | [-0.14; 0.15]  |
| 5   | { 100€, 0.5 ; 80€, 0.5 } | { 192.50€, 0.5 ; 5€, 0.5 } | -8.75€      | [0.15; 0.41]   |
| 6   | { 100€, 0.6 ; 80€, 0.4 } | { 192.50€, 0.6 ; 5€, 0.4 } | -25.50€     | [0.41; 0.68]   |
| 7   | { 100€, 0.7 ; 80€, 0.3 } | { 192.50€, 0.7 ; 5€, 0.3 } | -42.25€     | [0.68; 0.97]   |
| 8   | { 100€, 0.8 ; 80€, 0.2 } | { 192.50€, 0.8 ; 5€, 0.2 } | -59.00€     | [0.97; 1.37]   |
| 9   | { 100€, 0.9 ; 80€, 0.1 } | { 192.50€, 0.9 ; 5€, 0.1 } | -75.75€     | [1.37; $\infty$ )  |
| 10  | { 100€, 1.0 ; 80€, 0.0 } | { 192.50€, 1.0 ; 5€, 0.0 } | -92.50€     | non-monotone   |

Notes:  $E_A - E_B$  denotes the expected payoff difference between lottery A and lottery B. The payoffs above were used in the Internet experiment and the high stakes lab sessions. For the regular stakes lab sessions, the exchange rate into real money was 0.2. Participants were only shown the information of the first three columns. They had no explicit information on  $E$  or the CRRA coefficient.

have a relatively large difference in expected values, in this case 58.25€. As one proceeds down the list, the expected values of both lotteries increase, but eventually the expected value of lottery  $B$  exceeds the expected value of lottery  $A$ . Understanding HL tables is straightforward and explains their vast dispersion in experimental economics. The subject simply chooses lottery  $A$  or lottery  $B$  in each of the ten rows. After completion, one row is selected at random for payout of a certain subject. The chosen lottery is played out.

In terms of expected value, lottery  $B$  grows more attractive relative to lottery  $A$  with each step down the list. On the contrary, in each single row apart from row 10 lottery  $B$  bears the higher variance. The underlying logic of an HL table is that subjects face a tradeoff between expected value difference and variance. The last row is essentially a test whether the subject understood the instructions and has no relevance for risk aversion. It basically is the choice between 100€ and 192.50€ so every subject should choose Lottery  $B$  in row 10.<sup>18</sup> A risk neutral subject should choose the lottery with the higher expected value in each row. Hence a risk-neutral subject would choose  $A$  for the first four rows and  $B$  thereafter. We define the “switching point” as the first decision in which lottery  $B$  is chosen. A subject with an earlier switching point is classified as “risk loving” whereas a

<sup>18</sup>Monotonicity assumed.

subject with a later switching point is classified as “risk averse”.

To a rational agent, in row  $i$ , lottery  $B_i$  relative to lottery  $A_i$  should be at least as attractive as lottery  $B_{i-1}$  relative to lottery  $A_{i-1}$ . We should rarely see an agent choose  $B$  in one row and choose  $A$  in the next row. Andersen et al. (2006) and Harrison et al. (2007a) discuss conditions in which switching from  $B$  to  $A$  can occur without violating rationality. For example, simply a “fatter” indifference curve of a participant. We refrain from these rather special cases and label participants, who switch from  $B$  to  $A$  at any two consecutive rows or choose  $A$  in the ultimate row as “inconsistent”.<sup>19</sup>

Holt and Laury (2002) focus on their whole dataset including inconsistent participants. When building the graphs, they sort the participants’ choices. Meaning, all  $A$  choices are sorted to be before all  $B$  choices. This way, possible inconsistencies are assumed away.<sup>20</sup> This approach is debatable as it includes subjects who are potentially irrational or did not understand the experimental rules. In the best case this merely brings noise to the data, in the worst case, there are systematic biases. However, this approach entails the advantage of never having to worry about selection effects as the researcher is never forced to exclude subjects as “inconsistent”. One way to circumvent this calamity is to simply ask participants for their switching point instead of asking for their decision in each row. But this procedure imposes the single switching point on the participants. So the data will seem to be fine but are not generated in a way that fully reflects the participants desired choices. Our approach lets subjects choose the lottery for each row individually. In order to catch all effects, our analysis will regard the entire dataset as well as the dataset without inconsistent participants.

HL tables have the major drawback that their classification of “risk aversion” is not based on a general concept of risk. They require one certain utility framework (e.g. EUT) for all agents to classify their risk aversion relative to each other. There is increasing

---

<sup>19</sup>One could go further and distinguish between “inconsistency of type 1” for switchers from  $B$  to  $A$  and “inconsistency of type 2” for participants that chose the first-order stochastically dominated lottery  $A$  in the ultimate row. Arguably, the potentially underlying misconceptions of these two types are different. We do not think that this approach would add to our analysis and therefore refrain from it.

<sup>20</sup>It still may be the case that a subject chose  $A$  throughout the list.



literature on agents without an EUT compatible utility function, e.g. Harrison et al. (2010) and Harrison and Rutström (2009) find that agents seem to be split into groups using Prospect Theory or EUT or even a single person's choices are properly described by using both models in decision making. Maier and Rieger (2011) discuss this issue and propose a method to measure risk aversion model-independently via an MPL. Also, HL tables with their changing probabilities are sensitive to probability weighting of the participants. Another way to elicit risk aversion was proposed by Hey and Orme (1994) and is becoming increasingly popular. With the help of this approach, the researcher is able to get better estimates of the participants' risk aversion. This is mainly due to the fact that many observations (instead of just one from an MPL) are collected. In our design we needed an easy to implement task which would give us a fair estimate of a participant's risk aversion. In this regard, HL tables were best suited.

### 1.3.3 Inconsistency

The researcher has to regard inconsistent HL table choices in the analysis of risk aversion. This can help to distinguish decisions based on actual risk attitudes from mistakes.<sup>21</sup> This section deals with the influence of mistakes on measured risk aversion. In order to make a clear point, we treat the data like Holt and Laury (2002) in this example – we sort  $A$  and  $B$  choices and assume  $A$  choices to be in lower number rows than  $B$  choices.<sup>22</sup> Our main analysis later on does not sort the data in any way. For simplicity, we focus on a classical HL table with two lottery choices per row, ten decision rows, and a risk-neutral switching point from  $A$  to  $B$  (i.e. the first choice of  $B$ ) at row 5 just as in Table 1.1.<sup>23</sup> The argument can be generalized to MPLs with more than two lotteries per row, a different risk-neutral switching point, and a row number different from ten.

---

<sup>21</sup>Jacobson and Petrie (2009) discuss the added value of analyzing mistakes interacted with risk aversion data in a field experiment.

<sup>22</sup>The argument in this example is valid for all completed tables with a general tendency to choose  $A$  in lower number rows and  $B$  in higher number rows. This is the greatly predominant pattern found in completed HL tables of the usual design.

<sup>23</sup>A later switching point means risk aversion, an earlier one means risk love.

Suppose we have unbiased mistakes - choosing  $B$  instead of the true preference  $A$  or vice versa with equal probabilities and independent of the lottery probabilities/stakes. The more mistakes the MPL of an individual has, the stronger are our expectations of the answer distribution being shifted to an equal distribution of  $A$  and  $B$  choices. In the extreme case, an individual answers all rows randomly with an expected equal distribution of five  $A$  and five  $B$  choices – a switching point at row 6. In a regular HL table like ours, this implied a slightly risk averse individual. Would an HL table be designed with a switching point for a risk-neutral individual at a row number greater than 6 (at row 6), an MPL with only mistake choices would let us expectedly classify the individual as risk-loving (risk-neutral).

Generally said, mistakes gravitate the choice distribution towards equality. This argument is akin to the one in Harrison (1994). In HL graphs like in Figure 1.1, this means a line similar to the one of a risk neutral subject but with a switching point at decision 6, not at decision 5. Were the line of the true preferences to the southwest (northeast) of that line, we would get more (less) risk aversion or less (more) risk love via the mistakes. Whether it were more/less risk aversion/love depended on the risk-neutral switching point. Given the usual risk-neutral switching point at row 5 and the average pattern to choose  $A$  in more than five rows (risk averse behavior), inconsistency expectedly decreases risk aversion as a statistical artefact. As our data shows all usual patterns expected from HL data we anticipate such a result in our data as well.<sup>24</sup>

This effect has to be taken into account as it may bias the results of HL tables. Mistakes orthogonal to stake size and probabilities can result in biased data simply by the design of the table. Say we would find the usual pattern of higher risk aversion of women in our entire dataset. If the “female” dummy did not show a higher risk aversion coefficient in

---

<sup>24</sup>The whole argument is put forward without regarding probability weighting. Probability weighting participants would entail an increased frequency of inconsistent switching points closer to the inflection point of the probability weighting function. Depending on which weighting function was the more accurate one, the participants are more likely to be inconsistent around  $p = 0.5$  like in Quiggin’s (1982) function or around  $p = 1/e$  like in Prelec’s (1998) function. Regarding probability weighting correctly in our analysis would greatly increase the complexity and extent of our analysis – especially with our given dataset – without adding much quality to our results. We therefore do not go deeper into this subtopic.

the consistent dataset and women submitted less inconsistent HL tables than men, the risk aversion result would most probably be due to the inconsistencies but not to actual preferences. It is therefore essential to take a look at the whole dataset as well as at the consistent subset. The whole dataset will provide the researcher with maximum statistical power, while the consistent subset will help to pinpoint the results, especially concerning the demographic analysis. This is particularly useful for the comparison of Internet data (in which we expect more inconsistencies) to laboratory data. Results of the comparison from the entire dataset might vanish or even reverse if we analyze the consistent data only.

### 1.3.4 Random Incentive Schemes & Ambiguity Aversion

Ambiguity aversion could have influenced the risk aversion results of our online experiment. The literature on the relation between risk and ambiguity aversion shows mixed results. Cohen et al. (1987) as well as Curley et al. (1986) find ambiguity aversion to be uncorrelated to risk aversion. Potamites and Zhang (2007) find a slight positive correlation. Lauriola and Levin (2001) as well as Lauriola et al. (2007) find a positive correlation, just like Kocher and Trautmann (2013). Bossaerts et al. (2007) and Charness and Gneezy (2010) support this finding. While newer literature seems to endorse a positive correlation, overall there does not seem to be a clear consensus about this relationship. We cannot rule out that ambiguity aversion influenced the risk preferences of the online participants. However, the recruiting process of usual laboratory experiments brings along ambiguity as well. Participants are never told, how much money they can make and only frequent visitors can estimate a fair earnings interval. So we do not think that ambiguity massively influenced our results.

With our given Internet dataset, we cannot properly analyze ambiguity aversion of the participants let alone the correlation to risk aversion. However, we can get some suggestive results from our data. By definition, ambiguity aversion is negatively correlated with

participation in an experiment with an ambiguous payment scheme like in our Internet experiment. In case of a positive correlation of ambiguity aversion to risk aversion, participation should also be negatively correlated to risk aversion. This in turn would lead to a population of lower risk aversion than we would have gotten had we not introduced ambiguity in the payment procedure. In a classical HL graph, the curve would be shifted to the southwest. We briefly analyze this explanation in the concluding section.

## 1.4 Hypotheses

As a first hypotheses we think about control in two directions. As already stated, the experimenter has less control over the participant. She has less control over his understanding of the rules, his attention to and concentration on the task, and his attempts to game the experimental rules. Additionally, the participant himself also has less control over the experimenter. He cannot ensure his payment the way he can in a lab experiment or get questions answered in an extensive and personal way. Therefore, choices might get a hypothetical touch and/or choices might not be based on an actual understanding of the game. The former can make choices arbitrary, the latter can lead to noisy or even biased data. This hypothesis is analyzed by comparing the inconsistent behavior in the Internet and in the laboratory.

Our second and main hypothesis is that Internet experiments lead to data different from lab data. If true, this would entail different experimental conclusions than from laboratory experiments. This invalidated Internet experiments as a substitute for lab experiments. It is important to stress that we exclusively focus on the step of conducting a laboratory experiment with an RIS online. We do not analyze the step from the laboratory into the Internet as a technique to add field context. We analyze this hypothesis by comparing risk aversion between the Internet and the laboratory.

With the help of our first hypothesis we can split our main hypothesis into two more concrete hypotheses: (A) Data from the Internet is structurally different from lab data.

In our context this would mean a different appeal of a HL risk aversion graph. It may lead to flat line at a certain y-level, an ascending curve, or even a curve with descending as well as ascending sections. (B) Through more inconsistencies, the internet data shows less risk aversion than lab data with identical stakes.

## 1.5 Results

The descriptives and the nonparametric tests use the entire dataset. The analyses which include demographic data were conducted with the subgroup of participants which submitted the respective demographic information. This reduces the observation number of the respective analysis, but the remaining number of participants is still 2,400 and 2,700 for consistent and inconsistent participants, respectively. Statistical tests with the treatment levels only support our findings – although the results get sharpened by the demographics. We assume that submitting demographic data is orthogonal to risk and ambiguity aversion so that we do not regard selection as a reason for data confound. In this section of the analysis we do not sort the participants’ decisions in any way.

### 1.5.1 Descriptives

We collected MPL risk preference data from three different experimental treatments: the Internet treatment with an ambiguous selection probability and high stakes (see Table 1.1), one laboratory treatment with the same stakes and a choice selection probability of 0.667%, and another laboratory treatment with stake sizes divided by 5, but a 5 times higher choice selection probability of 3.333%.

Table 1.2 displays a summary of demographic characteristics of the participants. Our subject populations appear to be very similar across conditions for most demographic factors – a result of our efforts to mostly recruit for the Internet experiment from standard laboratory subject pools. This way we gave up an advantage of an internet experiment –

## INTERNET PAYMENT

Table 1.2: Demographic Characteristics of participants in the three experimental conditions

|                                   | Internet    | Lab High Stakes | Lab Low Stakes |
|-----------------------------------|-------------|-----------------|----------------|
| Number of participants            | 3,582       | 60              | 60             |
| Avg. response rate                | 96.8%       | 100.0%          | 100.0%         |
| Age <i>Avg. (StdDev)</i>          | 25.6 (5.8)  | 24.5 (4.2)      | 24.4 (3.9)     |
| Sex <i>male</i>                   | 62.5%       | 36.7%           | 36.7%          |
| Marital Status                    |             |                 |                |
| <i>Single</i>                     | 90.1%       | 94.9%           | 98.3%          |
| <i>Married</i>                    | 6.1%        | 5.1%            | 1.7%           |
| <i>Other</i>                      | 3.8%        | 0.0%            | 0.0%           |
| Charge of expenses                |             |                 |                |
| <i>Self</i>                       | 81.9%       | 81.7%           | 76.7%          |
| <i>Parents</i>                    | 11.3%       | 16.7%           | 20.0%          |
| <i>Spouse</i>                     | 1.1%        | 1.7%            | 1.7%           |
| <i>Other</i>                      | 5.6%        | 0.0%            | 1.7%           |
| Most frequent countries of origin |             |                 |                |
| <i>Germany</i>                    | 89.9%       | 85.0%           | 73.3%          |
| <i>China</i>                      | 1.0%        |                 |                |
| <i>Turkey</i>                     | 0.8%        |                 |                |
| <i>Russia</i>                     | 0.8%        | 3.3%            |                |
| <i>Poland</i>                     | 0.6%        |                 |                |
| <i>Bulgaria</i>                   |             | 3.3%            | 5.0%           |
| <i>Ukraine</i>                    |             | 3.3%            |                |
| <i>Kazakhstan</i>                 |             |                 | 3.3%           |
| <i>Other</i>                      | 7.1%        | 5.1%            | 18.4%          |
| Share of students                 | 79.8%       | 93.3%           | 96.7%          |
| For students:                     |             |                 |                |
| Semester <i>Avg. (StdDev)</i>     | 6.00 (3.75) | 4.55 (3.41)     | 5.11 (4.02)    |
| Field of Studies                  |             |                 |                |
| <i>Bus. Adm.</i>                  | 24.5%       | 43.9%           | 35.7%          |
| <i>Economics</i>                  | 18.8%       | 15.8%           | 16.1%          |
| <i>Engineering</i>                | 8.0%        | 1.8%            | 1.8%           |
| <i>Natural Science</i>            | 6.9%        | 3.5%            | 8.9%           |
| <i>Law</i>                        | 6.4%        | 5.3%            | 8.9%           |
| <i>Other</i>                      | 35.4%       | 29.8%           | 28.6%          |

the higher diversity of demographics.

An obvious difference is the higher share of male participants in the Internet treatment. Internet participants were recruited with explicit information of the underlying soccer markets. In contrast to that, laboratory participants were invited in the usual way without any announcement about the experimental topic and with exclusion of the online participants. Therefore, a lower share of women in the Internet treatment is expected.<sup>25</sup> We control for this by demographics coefficients in the econometric analysis. The general similarities in our subject pools and the demographic controls and our demographic

<sup>25</sup>In Europe, association football is a sport chiefly played and watched by men.

## INTERNET PAYMENT

controls let us be confident that any differences we might observe between our laboratory sessions and the Internet experiment are mainly due to the different environments and/or payment probabilities.

Figure 1.1: Proportion of safe choices in each decision

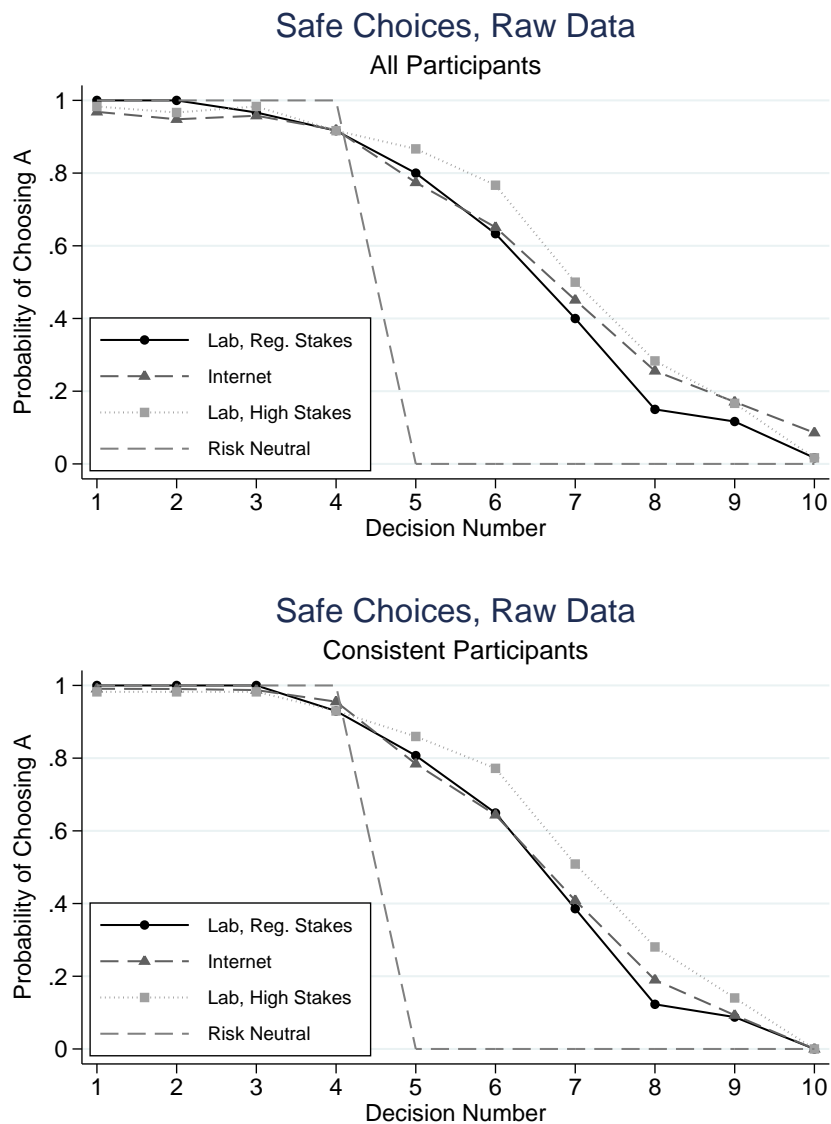


Figure 1.1 shows the share of *A* choices for each of the ten decisions in each of our three treatments, separately for the entire dataset (upper panel) and the subset of consistent participants (lower panel). As a first observation, the population as a whole of each treatment is risk-averse: compared to a risk-neutral population, there is a considerable

shift of the curve to the northeast in all treatments.<sup>26</sup> Comparing the treatments to each other, the Internet data crosses the lines of the other treatments and somewhat falls in between the curves of the regular and the high stake lab treatments. If anything, it is slightly closer to the lab regular stakes treatment. Narrowing our focus to consistent choices reveals a clearer picture: Internet participants behave very much like low stake lab participants, while high stake lab participants show a slightly stronger risk-aversion.

### 1.5.2 Inconsistency

We now turn to the number of consistent choices within each of our treatments. While in each of our laboratory conditions merely 5% of participants (3 out of 60 in both conditions, so 6 out of 120 overall) submit inconsistent choice profiles, the share of inconsistent participants in the Internet data is 14.3% (513 out of 3,582). We therefore expect a significant influence of the Internet environment on an inconsistent completion of our HL table.

We run a probit analysis on the whole dataset in order to find out, whether an Internet environment actually leads to inconsistency and what demographical factors have an influence on completing an MPL inconsistently. Table 1.3 summarizes the results. We can see that the probability to be inconsistent strongly increases if the experiment is conducted online instead of in a laboratory.

#### **Result 1 [Inconsistency, Treatment Comparison]**

*Internet participants show a strongly increased probability of inconsistent behavior compared to lab participants.*

There is no significant influence of the coefficient for the lab high stakes treatment. This leads to the conclusion that there is no effect of the stake size on filling out the question-

---

<sup>26</sup>Like in all HL graphs, a curve northeast to the risk neutral curve means risk aversion and a curve southwest to the risk neutral curve means risk love.



INTERNET PAYMENT

Table 1.3: Probit Estimates of Inconsistent Choice

|                                   | Model 1     |           | Model 2     |           |
|-----------------------------------|-------------|-----------|-------------|-----------|
|                                   | Coefficient | Std. Err. | Coefficient | Std. Err. |
| Constant                          | -2.0807***  | (0.3008)  | -2.5919***  | (0.4417)  |
| Internet                          | 0.6542**    | (0.2791)  | 0.8514***   | (0.3284)  |
| Lab High Stakes                   | 0.0169      | (0.3896)  | 0.2540      | (0.4346)  |
| Female                            | 0.2448***   | (0.0532)  | 0.2592***   | (0.0638)  |
| Age                               | 0.0103**    | (0.0042)  |             |           |
| Age (at begin of studies)         |             |           | 0.0202*     | (0.0123)  |
| Semester                          |             |           | -0.0244***  | (0.0089)  |
| <i>Field of studies</i>           |             |           |             |           |
| Economics                         |             |           | -0.0682     | (0.1033)  |
| Engineering                       |             |           | 0.1867      | (0.1293)  |
| Natural Science                   |             |           | 0.1635      | (0.1376)  |
| Law                               |             |           | 0.4277***   | (0.1280)  |
| Other                             |             |           | 0.308***    | (0.0817)  |
| <i>Home country</i>               |             |           |             |           |
| China                             |             |           | 0.1406      | (0.3064)  |
| Turkey                            |             |           | 0.5486*     | (0.2900)  |
| Russia                            |             |           | 0.2658      | (0.2928)  |
| Poland                            |             |           | -0.1099     | (0.3964)  |
| Bulgaria                          |             |           | 0.1693      | (0.5686)  |
| Ukraine                           |             |           | 0.3658      | (0.3941)  |
| Kazakhstan                        |             |           | (omitted)   |           |
| Other                             |             |           | 0.6442***   | (0.1144)  |
| <i>Charge of expenses</i>         |             |           |             |           |
| Parents                           |             |           | 0.0897      | (0.0912)  |
| Spouse                            |             |           | 0.3771      | (0.2662)  |
| Other                             |             |           | -0.0761     | (0.1536)  |
| <i>Marital status</i>             |             |           |             |           |
| Married                           |             |           | 0.2127      | (0.2032)  |
| Other                             |             |           | 0.2565      | (0.1624)  |
| Number of obs                     | 3,594       |           | 2,832       |           |
| Log likelihood                    | -1,446      |           | -1,059      |           |
| LR $\chi^2(4)$ resp. $\chi^2(22)$ | 36.11       |           | 122.35      |           |
| Prob > $\chi^2$                   | 0.0000      |           | 0.0000      |           |
| Pseudo $R^2$                      | 0.0123      |           | 0.0546      |           |

Notes: \*, \*\*, and \*\*\* indicate significance at the 10%, 5% and 1% level, respectively. The baseline category is “Lab Low Stakes”. In Model 2, omitted variables are the ones with the highest frequency in their respective class: Business Administration (field of studies), Germany (home country), Self (charge of expenses), Single (marital status). Additionally, Home country Kazakhstan is omitted because of perfect prediction (missing variance).

naire inconsistently. So it does not seem that participants of our experiment had to be financially motivated to take the experiment seriously. The regular lab stakes sufficed.<sup>27</sup> Female subjects show a significantly increased probability of inconsistency. The participants' age at the beginning of studies as well as semester also have a (marginally) significant influence on the probability of inconsistency. The influence goes in opposite - albeit expectable - directions. Subjects, who were older at the beginning of their studies have an increased probability of inconsistency. We expect them to have repeated at least one year in school (which in the German school system is implemented if the pupil performed insufficiently that very year) or qualified for university studies after having already worked in a job which required a lesser degree<sup>28</sup>. So we expected a higher probability of inconsistency for these participants. A higher semester is associated with a higher education. This made us expect a lower probability for inconsistency. However, the actual impact of the semester and age at beginning of studies coefficients is economically negligible. We do not find any influences of marital status or who is in charge of the expenses of the household.

Students of law and “other” majors are much more likely to be inconsistent than students of the other categories. An anticipated result, as both categories of studies are the ones with the least math classes. (“Other” contained the humanities.) Country dummies, which have a significant effect are Turkey and “Other”. The latter includes countries like Colombia, France, Hungary, Italy, the Netherlands, the USA, and Vietnam. We cannot find any clear cultural pattern in “Other” countries. So overall the tendency to inconsistency seems to rather spring from language problems than from one clear cultural difference to German participants. It is fair to assume this as the reason for the increased inconsistency probability of Turkish participants as well.

## **Result 2 [Inconsistency, Demographic Analysis]**

---

<sup>27</sup>Alternatively, the participants computed the task like one big compound lottery, including the exchange rates of the stakes. In this case, there was no expected stake size difference in the two lab treatments. Participants do not seem to have done this, see Section 1.2 and Section 1.5.3.

<sup>28</sup>This is called “zweiter Bildungsweg” – “second way of education” – in Germany. This includes all ways to earn a qualification to enter university apart from the regular school path.

(a) *The stake size and the lottery procedure do not influence the probability of inconsistent behavior in the laboratory.*

(b) *Female subjects as well as subjects which were older at the beginning of their studies have an increased probability of inconsistent behavior. Subjects in later semesters show a decreased probability.*

(c) *Untrained math skills as well as language problems increase the probability of inconsistent behavior.*

### 1.5.3 Risk Aversion

From here on we focus on the risk aversion analysis. We choose three methods in order to analyze the data. The first part relies on nonparametric tests. This is followed by ordered probit analysis, our main tool. Both approaches have the advantage that they are agnostic about functional forms of the utility function. The ordered probit analysis additionally offers to employ control variables. The final part employs an interval regression. A CRRA utility function is assumed and we run regressions with intervals for the risk aversion parameter.

#### Nonparametric Tests

To nonparametrically test for differences in risk attitudes, we conduct Kolmogorov-Smirnov (KS) tests comparing the distribution of safe choice frequencies between treatments. Table 1.4 lists the results. The differences we observe between our treatments are not significant, neither for comparing our lab treatments to each other nor for comparing Internet data with each of our lab treatments. We obtain this result for the entire dataset as well as for the consistent data only. So as our first tentative result we have three payment schemes which lead to similar risk aversion measures in HL tables. A closer look at the data might reveal slight differences or differences based on variables we did not control for in the nonparametric approach.

Table 1.4: Results of Kolmogorov-Smirnov tests on differences in distribution of safe choices

|                      | All choices | Consistent choices |
|----------------------|-------------|--------------------|
|                      | p-value     | p-value            |
| Lab (low stakes) vs. |             |                    |
| Internet             | 0.732       | 0.947              |
| Lab (high stakes)    | 0.432       | 0.398              |
| Internet vs.         |             |                    |
| Lab (high stakes)    | 0.240       | 0.262              |

Note: All tests are two-sided.

The nonparametric tests and also Figure 1.1 show that our Internet data is not structurally different from our lab data.

### Result 3 [Risk Aversion, Structural Data Differences]

*The Internet as an experimental environment does not lead to structurally different data compared to laboratory data.*

### Ordered Probit Regressions

As the main method we employ ordered probit regressions. For analyzing HL tables they were introduced by Harrison et al. (2005) and we essentially use their publicly available Stata code.<sup>29</sup> They put forward the argument that ordered probit regressions do not rely on any functional form of the utility function and take care of the order as well as the interdependencies of the ten decisions in an HL table. Therefore it is an appropriate method to analyze this kind of data.

The given Stata code does not count the safe choices like Holt and Laury (2002) do in their analysis. It instead uses the minimum switching point (MSP) as the dependent variable. The MSP is the first decision in which a subject chooses Lottery  $B$ . This way, the first choice of the “risky” lottery is marked as the pivotal decision in this context. Switching back and forth after this decision is not included in the analysis.<sup>30</sup> In the case

<sup>29</sup>We kindly thank Harrison, Johnson, McInnes, and Rutström for making the code accessible on the ExLab website.

<sup>30</sup>Which is the difference of the “minimum switching point” to the “switching point”. The latter can only be used for consistent data.

of such a “fatter” indifference curve, only the most risk loving border of the indifference curve is regarded through the MSP.<sup>31</sup> We run the whole analysis with the MSP as the dependent variable (like Harrison et al. (2005)) as well as with the number of *A* choices (in the spirit of Holt and Laury (2002)) as the dependent variable in order to check for the results’ robustness. Note, that this code cannot deal with participants, who chose Option *A* throughout the whole list. Therefore, these 176 participants (1 of them being a lab participant) are excluded from the ordered probit analysis.<sup>32</sup>

Figure 1.2 (upper panel) is obtained by running an ordered probit regression model with an extensive set of control variables<sup>33</sup> on the whole dataset. The first observation of the graph is that the curve of the Internet data is virtually identical to the graph of the regular lab stakes data.<sup>34</sup> The second observation is the line of the lab high stakes data being stronger curved to the northeast than the other lines and all its datapoints apart from the one at decision 10 are lying above the datapoints of the other lines. Despite this substantial graphic difference the effect is not significant in the ordered probit regression. The picture becomes clearer if we only regard the consistent data in the lower panel of Figure 1.2. The regular lab treatment now shows the least risk aversion. The internet treatment shows a slightly stronger risk aversion, although not significantly so in the regression. Again, the lab high stakes treatment graphically shows most risk aversion, this is also significant in the regression.

Table 1.5 shows the regression results. When comparing Model 2 (Consistent Data) to Model 1 (All Data), we can see that the propensity to submit an inconsistent HL table interacts with risk aversion. In the category “Field of Studies” both Law and “Other”

---

<sup>31</sup>Alternative approaches include using the most risk averse border or the middle of the curve. In some sense, the expected effect of inconsistencies is maximized this way. There are about six *A* choices of a representative individual. So there is a higher probability for a *B* choice at an unexpected decision row than for an *A* choice.

<sup>32</sup>They make for 4.9% of the Internet participants and 0.83% of the laboratory participants. We can see that the percentage of A-only participants relative to all inconsistent participants doubles when comparing the lab (16.67%) to the Internet (34.11%). This result is most probably due to misunderstanding the experimental rules and can be interpreted as further support for the claim that the experimenter loses control in the Internet.

<sup>33</sup>The same set like in the probit analysis of inconsistent choices.

<sup>34</sup>There are differences, but they are never greater than 0.0005.

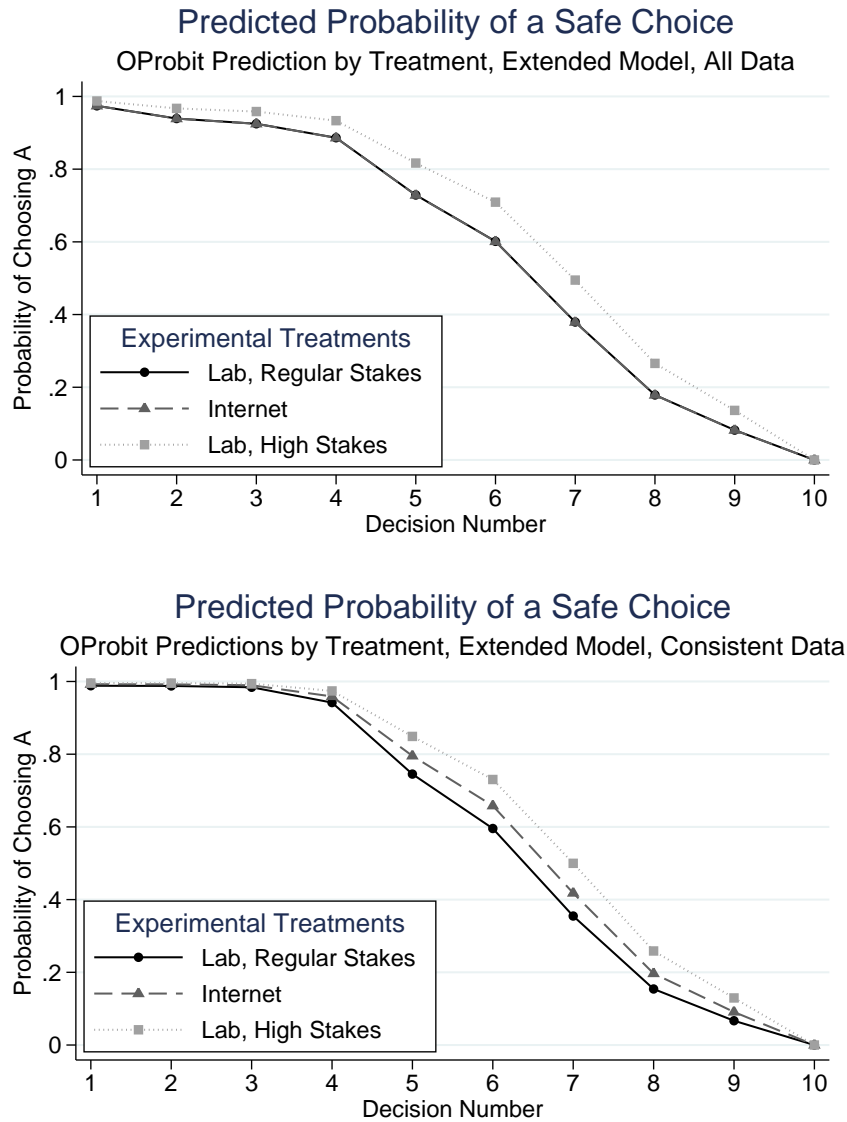
INTERNET PAYMENT

Table 1.5: Ordered probit estimates (MSP) of likelihood to choose lottery A - models with full demographics

|                           | Model 1     |           | Model 2         |           |
|---------------------------|-------------|-----------|-----------------|-----------|
|                           | All Data    |           | Consistent Data |           |
|                           | Coefficient | Std. Err. | Coefficient     | Std. Err. |
| Internet                  | -0.0013     | (0.1399)  | 0.1668          | (0.1435)  |
| Lab High Stakes           | 0.2957      | (0.1953)  | 0.3746*         | (0.2016)  |
| Female                    | -0.0190     | (0.0420)  | 0.0501          | (0.0447)  |
| Age (at begin of studies) | -0.0090     | (0.0081)  | -0.0035         | (0.0089)  |
| Semester                  | 0.0111**    | (0.0054)  | 0.0027          | (0.0056)  |
| <i>Field of studies</i>   |             |           |                 |           |
| Economics                 | 0.0884      | (0.0599)  | 0.0749          | (0.0619)  |
| Engineering               | 0.0576      | (0.0807)  | 0.1215          | (0.0848)  |
| Natural Science           | 0.2958***   | (0.0807)  | 0.3443***       | (0.0902)  |
| Law                       | -0.0136     | (0.0884)  | 0.2084**        | (0.0959)  |
| Other                     | 0.0453      | (0.0516)  | 0.1832***       | (0.0544)  |
| <i>Home country</i>       |             |           |                 |           |
| China                     | -0.0299     | (0.2063)  | 0.0609          | (0.2213)  |
| Turkey                    | -0.1926     | (0.2143)  | 0.0977          | (0.2433)  |
| Russia                    | 0.1079      | (0.2209)  | 0.1964          | (0.2390)  |
| Poland                    | 0.0844      | (0.2297)  | 0.1716          | (0.2438)  |
| Bulgaria                  | 0.3981      | (0.3461)  | 0.4620          | (0.3687)  |
| Ukraine                   | 0.1036      | (0.2869)  | 0.4918          | (0.3297)  |
| Kazakhstan                | 0.1955      | (0.3233)  | 0.1205          | (0.3242)  |
| Other                     | -0.1577*    | (0.0891)  | 0.1130          | (0.1008)  |
| <i>Charge of expenses</i> |             |           |                 |           |
| Parents                   | 0.0198      | (0.0602)  | 0.0685          | (0.0639)  |
| Spouse                    | -0.0874     | (0.2056)  | -0.0117         | (0.2281)  |
| Other                     | 0.0747      | (0.0979)  | 0.1420          | (0.1050)  |
| <i>Marital status</i>     |             |           |                 |           |
| Married                   | -0.1101     | (0.1468)  | -0.0671         | (0.1596)  |
| Other                     | -0.2070*    | (0.1157)  | 0.0214          | (0.1281)  |
| cut1                      | -2.0472     | (0.2412)  | -2.1774         | (0.2649)  |
| cut2                      | -1.6460     | (0.2382)  | -2.1579         | (0.2643)  |
| cut3                      | -1.5364     | (0.2377)  | -2.0567         | (0.2615)  |
| cut4                      | -1.3036     | (0.2369)  | -1.4752         | (0.2543)  |
| cut5                      | -0.7018     | (0.2360)  | -0.5566         | (0.2520)  |
| cut6                      | -0.3474     | (0.2357)  | -0.1360         | (0.2517)  |
| cut7                      | 0.2201      | (0.2358)  | 0.4830          | (0.2518)  |
| cut8                      | 0.8370      | (0.2364)  | 1.1346          | (0.2525)  |
| cut9                      | 1.3107      | (0.2375)  | 1.6198          | (0.2536)  |
| Log likelihood            | -5,503      |           | -4,561          |           |
| Number of obs             | 2,709       |           | 2,460           |           |
| LR $\chi^2(23)$           | 36.77       |           | 37.82           |           |
| Prob > $\chi^2$           | 0.0343      |           | 0.0267          |           |
| Pseudo $R^2$              | 0.0033      |           | 0.0041          |           |

Notes: \*, \*\*, and \*\*\* indicate significance at the 10%, 5% and 1% level, respectively. The baseline category is “Laboratory”. Omitted variables are the ones with the highest frequency in their respective class: Business Administration (field of studies), Germany (home country), Self (charge of expenses), Single (marital status). In Model 2, only data from consistent choice profiles are included.

Figure 1.2: Predicted safe choices



turn from an insignificant to an increased and significantly positive coefficient, meaning higher risk aversion. Both fields of studies had a high and significant increased probability to submit an inconsistent HL table. An analogue effect can be found for the coefficient of “Other” home country. It turns from being significantly negative to insignificantly positive. This category also had a highly increased probability of submitting an inconsistent HL table. So we can see that inconsistency had the effect we described in Section refs:rd:inconsistency – it reduces the measured risk aversion.

When focusing on the number of safe choices as the dependent variable in our analysis, our main results are essentially identical to the analysis of the MSP. Furthermore, the increased risk aversion of the laboratory high stakes data is already significant in the regression results of the entire dataset. This robustness check lets us be confident of the quality of our results. All results of the ordered probit analysis for the number of safe choices can be found in the appendix.

Summarizing the ordered probit analysis, we do not find meaningful differences between the regular stakes lab treatment and the Internet treatment. Significant differences in risk aversion can be found for the lab high stakes data. This can be seen graphically as well as in the magnitude and the significance of the coefficients in the regressions. As a main message we can say that there seems to be no difference in risk aversion between the regular stakes lab treatment and the Internet treatment. However, lab subjects in the high stakes treatment behave in accordance to expectations more risk aversely.

Therefore, the ordered probit analysis strengthens our main results of the nonparametric tests. This analysis also gives us more information in the sense that they detect the higher risk aversion of the laboratory high stakes participants.

### Interval Regressions

As a final test we run interval regressions. The dependent variable is the risk aversion coefficient  $r$  in this constant relative risk aversion (CRRA) utility function:

$$U(x) = \begin{cases} \frac{x^{(1-r)}}{1-r} & \text{for } r \neq 1 \\ \ln x & \text{for } r = 1 \end{cases} \quad (1.1)$$

An  $r > 0$  represents risk aversion, an  $r = 0$  risk neutrality, and an  $r < 0$  risk love. So the greater the regression coefficient the greater the risk aversion (the smaller the risk love), and the smaller the coefficient the smaller the risk aversion (the greater the risk love). We use the intervals of the CRRA coefficient for each participant, displayed in the last column of Table 1.1. In order to use all data and not only the consistent data, we allow



the intervals to be broader than the ones in the table. The choice which determines the minimal  $r$  coefficient is the first chosen  $B$  lottery and the choice which determines the maximal  $r$  is the last chosen  $A$  lottery. This way we can regard subjects that switched from  $B$  to  $A$ , switched more than once, and even subjects who chose  $A$  in each row.

A subject, who chose  $A$  in the first two rows,  $B$  in the third row, again  $A$  in the fourth row, and finally  $B$  in all subsequent rows would be assigned to the interval  $[-0.95; 0.15]$ . The same subject, however, would be excluded from the analysis of consistent participants. A subject with  $B$  choices throughout would have a coefficient in the interval  $(-\infty; -1.71]$ . A subject with  $A$  choices throughout is assigned a coefficient in the interval  $[1.37; \infty)$  just like a subject with  $A$  choices in rows 1 to 9 and a  $B$  choice in the last row. Although this is formally not correct, we employ this approach in order to use all data in the first model. As usual, we also run the whole analysis with consistent data only (and thereby do not use the pure  $A$  choice participants).

In this approach we begin to put structure on the data. We assume EUT and the specified CRRA utility function to be valid for every participant. As mentioned before, an increasing body of literature is showing that people cannot easily be assigned to a certain utility function. Often enough, a population is split into groups of different classes as in Harrison et al. (2010) or even one individual behaves according to different functions in different situations as in Harrison and Rutström (2009). With the help of the intervals (of flexible size) we leave wiggle room for the data to fit into the regression parameters and let the data speak for itself. Par for the course of the stronger assumptions, the results are not as strong as our previous results but still in line with them. In this context, it makes sense to have a closer look at the consistent data. This data is closer to strict EUT-generated data than the entire dataset. Thereby, we reduce the number of datapoints that are not strictly in line with EUT and more probably derived from a EUT-incompatible utility function. Table 1.6 summarizes the results.

In the whole dataset, laboratory high stakes has a positive regression coefficient indicating an increased risk aversion. Internet - in contrast to our previous findings - has a substantial

INTERNET PAYMENT

Table 1.6: Interval regression estimates of CRRA coefficient - models with full demographics

|                           | Model 1     |           | Model 2         |           |
|---------------------------|-------------|-----------|-----------------|-----------|
|                           | All Data    |           | Consistent Data |           |
|                           | Coefficient | Std. Err. | Coefficient     | Std. Err. |
| Constant                  | 0.4826***   | (0.1185)  | 0.5113***       | (0.1330)  |
| Internet                  | 0.1298      | (0.0807)  | 0.0775          | (0.0760)  |
| Lab High Stakes           | 0.1689      | (0.1136)  | 0.1772*         | (0.1068)  |
| Female                    | 0.0453*     | (0.0245)  | 0.0309          | (0.0237)  |
| Age (at begin of studies) | -0.0027     | (0.0048)  | -0.0038         | (0.0047)  |
| Semester                  | 0.0000      | (0.0031)  | 0.0011          | (0.0030)  |
| <i>Field of studies</i>   |             |           |                 |           |
| Economics                 | 0.0344      | (0.0344)  | 0.0281          | (0.0328)  |
| Engineering               | 0.0755      | (0.0471)  | 0.0596          | (0.0449)  |
| Natural Science           | 0.1850***   | (0.0496)  | 0.1759***       | (0.0477)  |
| Law                       | 0.1229**    | (0.0525)  | 0.1080**        | (0.0508)  |
| Other                     | 0.0946***   | (0.0300)  | 0.0893***       | (0.0288)  |
| <i>Home country</i>       |             |           |                 |           |
| China                     | -0.0407     | (0.1215)  | -0.0153         | (0.1173)  |
| Turkey                    | -0.0160     | (0.1267)  | 0.0521          | (0.1287)  |
| Russia                    | 0.1390      | (0.1269)  | 0.0590          | (0.1266)  |
| Poland                    | 0.0337      | (0.1383)  | 0.0873          | (0.1293)  |
| Bulgaria                  | 0.2116      | (0.2055)  | 0.2187          | (0.1951)  |
| Ukraine                   | 0.1721      | (0.1780)  | 0.2811          | (0.1757)  |
| Kazakhstan                | 0.0048      | (0.1859)  | 0.0480          | (0.1724)  |
| Other                     | 0.1089      | (0.0528)  | 0.0596          | (0.0533)  |
| <i>Charge of expenses</i> |             |           |                 |           |
| Parents                   | 0.0370      | (0.0351)  | 0.0300          | (0.0339)  |
| Spouse                    | 0.0492      | (0.1205)  | 0.0082          | (0.1209)  |
| Other                     | 0.0496      | (0.0585)  | 0.0861          | (0.0557)  |
| <i>Marital status</i>     |             |           |                 |           |
| Married                   | -0.0071     | (0.0858)  | -0.0438         | (0.0846)  |
| Other                     | -0.0225     | (0.0708)  | -0.0031         | (0.0678)  |
| $\ln(\sigma)$             | -0.5534***  | (0.0157)  | -0.6312***      | (0.0158)  |
| $\sigma$                  | 0.5750      | (0.0090)  | 0.5319          | (0.0084)  |
| Log likelihood            | -5,176      |           | -4,799          |           |
| Number of obs             | 2,803       |           | 2,460           |           |
| LR $\chi^2(23)$           | 39.67       |           | 37.58           |           |
| Prob > $\chi^2$           | 0.0167      |           | 0.0283          |           |
| left-censored obs         | 31          |           | 18              |           |
| uncensored obs            | 0           |           | 0               |           |
| right-censored obs        | 476         |           | 225             |           |
| interval obs              | 2,296       |           | 2,217           |           |

Notes: \*, \*\*, and \*\*\* indicate significance at the 10%, 5% and 1% level, respectively. The baseline category is “Laboratory”. Omitted variables are the ones with the highest frequency in their respective class: Business Administration (field of studies), Germany (home country), Self (charge of expenses), Single (marital status). In Model 2, only data from consistent choice profiles are included.

positive coefficient. However, both coefficients are not significantly different from 0. This finding changes if we take a look at the consistent data only. The coefficient of Internet is reduced by 50% and stays insignificant. To the contrary, the coefficient of lab high stakes is slightly increased and now significantly greater than 0. In both datasets the dummies for natural sciences, law, and “Other” categories of studies show a significant increase in risk aversion compared to the omitted category (business administration). Similar effects have been found in the ordered probit analysis (for both switching point and number of safe choices). The interval regression analysis of the whole dataset is the only analysis in which we find the usual effect of women being more risk averse than men. We cannot find this effect in the ordered probit analysis (MSP and safe choices) for the whole and the consistent dataset and also not in the interval regression for the consistent subset of data. Overall, we do not find support in our data for risk attitudes differing by sex, just like Harrison et al. (2007b).<sup>35</sup>

Overall, our results lead us to a similarity of risk aversion in the lab regular stakes treatment and the Internet treatment. This is shown in nonparametric tests, in the ordered probit analysis, and in the interval regression analysis. The lab high stakes data shows a higher risk aversion than the lab regular stakes data. Graphically, the lab high stakes data also shows a higher risk aversion than the Internet data. This is supported by a post-estimation test in the MSP ordered probit analysis of the entire dataset. However, this result cannot be replicated in post-estimation tests of the ordered probit regression of the consistent data, the number of safe choices ordered probit regression (entire dataset & consistent dataset), or the interval regression analysis. The pattern of the test results hints to the more prevalent inconsistencies in the Internet data as a reason.

#### **Result 4 [Risk Aversion]**

*(a) Our Internet data does not show a different level of risk aversion than the laboratory regular stakes data does. Neither for the entire dataset nor for the consistent-only dataset.*

---

<sup>35</sup>For an extensive overview on conditions under which this pattern shows up, see Eckel and Grossman (2008).

(b) *The laboratory high stakes data shows significantly higher risk aversion.*

### **Result 5 [Risk Aversion]**

*Compared to the lab high stakes treatment, our Internet data does not show lower risk aversion.*

### **Experimental Costs**

Of practical interest to all experimenters are the costs relative to the quality and quantity of the data. Does it really make sense for the researcher to leave the lab? Is the loss of control overcompensated by more observations? This part of the analysis relates the costs and the number of participants to each other. In this analysis we focus on expected participant payments and assume that the paying researcher is also only interested in *expected* costs of the experiment.

Most important for us is the comparison of the Internet and the regular lab stakes data. In the Internet treatment we gathered 60 times as many participants as in the the lab regular stakes treatment (3,582 vs. 60). Focusing on the consistent participants, we gathered 54 times as many as in the lab regular stakes treatment (3,069 vs. 57). So the remarkable increase in participant numbers clearly overcompensated via statistical power (and broader demographics) for the increase in costs, whether they are monetary (we paid more participants in the lab than online) or in the form of data quality (we found more inconsistent behavior online).<sup>36</sup>

### **Result 6 [Experimental Costs]**

*The Internet provides the researcher with a greatly increased number of participants.*

*Potential higher monetary costs for higher stakes or paying more participants or costs in*

---

<sup>36</sup>Please note that we assume a not too strongly decreasing marginal utility in the number of observations for the researcher. The curvature of the researcher's utility function in the number of participants is dependent on the method of analysis, the researcher's preferences etc. and can be configured or is well-known by the researcher in advance of the experiment. We also abstain from including the recruitment and programming costs into our analysis as they are as well highly dependent on the specific researcher's options and preferences.

*terms of data quality loss are overcompensated by remarkably higher statistical power.*

## 1.6 Conclusions

As a general result we do not find different elicited risk attitudes in our regular stakes laboratory treatment and the Internet. We would have drawn the same conclusions regardless whether we conducted the experiment in the lab or in our online setting. When we also take a look at the high stakes laboratory treatment, we observe more risk aversion than in the regular stakes lab and Internet treatments. This is consistent with standard theory and the result one would expect according to the literature discussed in Section 1.2.1. When comparing the lab high stakes treatment to the Internet treatment (which have identical stakes) we see that the risk aversion in the Internet is lower than in the laboratory. Possible explanations include that Internet participants discount the stakes in some way or that a selection bias (through a positive correlation to ambiguity aversion) lets relatively more risk tolerant participants enter the experiment.

The latter point is important to stress. We do not obtain similar results for lab high stakes and Internet data although they employ identical stake sizes. The researcher has to adjust the stake sizes accordingly when leaving the lab. To what magnitude “accordingly” translates in such an environment (are our five times scaling up always the best idea?) and which factors influence this magnitude (ambiguity, payment insecurity, opportunity costs etc.) is left for future research.

We find more inconsistently completed HL tables online than in the laboratory. But we can also show that the increased inconsistency frequency is overcompensated by a strongly increased number of participants. The Internet helps to motivate more people to participate than a comparable lab experiment which has an upper bound of participants through time and lab space.

## INTERNET PAYMENT

So we conclude that it is feasible to run lab kind experiments online, if the researcher increases the stake size. Nevertheless, Internet experiments still lack the degree of control a lab experiment can offer. This is verified in generally noisier data and less risk aversion compared to similar stakes in the laboratory. Laboratory experiments are still the gold standard in terms of control. Future research has to explore the extent to which our results are applicable to other experimental research domains. More complex individual decision problems are a natural choice.

## 1.7 Appendix

### 1.7.1 Screenshots

Figure 1.3: Screenshot of the Registration Page

WORLD CUP 2006  
TORLABOR SOCCER TRADING MARKETS

**Registration**

To register for the TorLabor markets please fill in the form below. You will receive a confirmation e-mail upon our receipt of your completed form. This e-mail contains a link which you must follow to complete the registration process.

First name:

Last name:

E-mail:

Repeat e-mail:

Language: English

For each match of the World Cup 2006, a new market will open a few days before the start of the match. **Would you like to be informed via e-mail each time a new market opens?**

Yes.  No.  
(It is always possible to enable/disable the delivery of e-mails.)

Please choose a username and a password. Both are necessary to participate in the markets. For your username, use only letters, numbers, dot (".") or underscore ("\_"). Your password must be at least 6 characters long.

Username:

Password:

Repeat password:

Please read carefully the [rules](#) and the [terms and conditions](#) (T&Cs). If after reading the rules and the terms and conditions, you understand and accept them, check the following box:

I have read the rules and the terms and conditions and I accept them.

**Important:** Multiple registrations of the same person are NOT permitted and will lead to exclusion from participation and payment.

# INTERNET PAYMENT

Figure 1.4: Screenshot of the Risk Aversion Elicitation Page

WORLD CUP 2006  
 TORLABOR SOCCER TRADING MARKETS

**Note:** Thank you for confirming your registration. Please read carefully and follow the instructions below.

### Instructions

The bottom of the screen shows ten decisions. Each decision is a paired choice between "Option A" and "Option B." You are asked to make ten choices, and by completing this task you are offered a first possibility to earn some money at TorLabor soccer trading markets.

Later on, you will be offered more possibilities to earn money at TorLabor soccer trading markets. In particular, you may earn money by trading at TorLabor markets. Further details will be provided in due time. Please notice that if you are rewarded for completing the present task then you won't be rewarded for trading and that this information will be provided to you only after the World Cup 2006.

**How your choices affect your payoffs**

At the end of the World Cup 2006, 5 participants will be randomly selected. Imagine that you are one of these 5 randomly selected participants. A ten-sided die will be used to determine your payoffs. The faces of the die are numbered from 1 to 10. The ten-sided die will be thrown twice. The first throw will determine which of the 10 decisions you made will affect your payoffs. For this decision we will look at the option you have chosen. Then, for the chosen option, the die will be thrown a second time. You will receive the payoffs attached to the number of the die throw's result.

Thus, even though you will make ten decisions, only one of these will end up affecting your payoffs. You will not know in advance which decision will be used. Each decision has an equal chance of being used in the end.

Now, please look at Decision 1 at the top. Option A pays 100.00 euros if the throw of the ten-sided die yields 1, and it pays 80.00 euros if the throw yields 2, 3, 4, 5, 6, 7, 8, 9, or 10. Option B pays 192.50 euros if the throw of the die yields 1, and it pays 5.00 euros if the throw yields 2, 3, 4, 5, 6, 7, 8, 9, or 10. The other decisions are similar, except that as you move down the table, the chances of the higher payoff for each option increase. In fact, for Decision 10 in the bottom row, the die will not be needed since each option pays the highest payoff for sure, so your choice here is between 100 euros or 192.50 euros.

| Decision | Option A                |              | I choose option A     | Option B                |              | I choose option B     |
|----------|-------------------------|--------------|-----------------------|-------------------------|--------------|-----------------------|
|          | Die throw yields        | Option pays  |                       | Die throw yields        | Option pays  |                       |
| 1        | 1                       | 100.00 euros | <input type="radio"/> | 1                       | 192.50 euros | <input type="radio"/> |
|          | 2,3,4,5,6,7,8,9 or 10   | 80.00 euros  |                       | 2,3,4,5,6,7,8,9 or 10   | 5.00 euros   |                       |
| 2        | 1 or 2                  | 100.00 euros | <input type="radio"/> | 1 or 2                  | 192.50 euros | <input type="radio"/> |
|          | 3,4,5,6,7,8,9 or 10     | 80.00 euros  |                       | 3,4,5,6,7,8,9 or 10     | 5.00 euros   |                       |
| 3        | 1,2 or 3                | 100.00 euros | <input type="radio"/> | 1,2 or 3                | 192.50 euros | <input type="radio"/> |
|          | 4,5,6,7,8,9 or 10       | 80.00 euros  |                       | 4,5,6,7,8,9 or 10       | 5.00 euros   |                       |
| 4        | 1,2,3 or 4              | 100.00 euros | <input type="radio"/> | 1,2,3 or 4              | 192.50 euros | <input type="radio"/> |
|          | 5,6,7,8,9 or 10         | 80.00 euros  |                       | 5,6,7,8,9 or 10         | 5.00 euros   |                       |
| 5        | 1,2,3,4 or 5            | 100.00 euros | <input type="radio"/> | 1,2,3,4 or 5            | 192.50 euros | <input type="radio"/> |
|          | 6,7,8,9 or 10           | 80.00 euros  |                       | 6,7,8,9 or 10           | 5.00 euros   |                       |
| 6        | 1,2,3,4,5 or 6          | 100.00 euros | <input type="radio"/> | 1,2,3,4,5 or 6          | 192.50 euros | <input type="radio"/> |
|          | 7,8,9 or 10             | 80.00 euros  |                       | 7,8,9 or 10             | 5.00 euros   |                       |
| 7        | 1,2,3,4,5,6 or 7        | 100.00 euros | <input type="radio"/> | 1,2,3,4,5,6 or 7        | 192.50 euros | <input type="radio"/> |
|          | 8,9 or 10               | 80.00 euros  |                       | 8,9 or 10               | 5.00 euros   |                       |
| 8        | 1,2,3,4,5,6,7 or 8      | 100.00 euros | <input type="radio"/> | 1,2,3,4,5,6,7 or 8      | 192.50 euros | <input type="radio"/> |
|          | 9 or 10                 | 80.00 euros  |                       | 9 or 10                 | 5.00 euros   |                       |
| 9        | 1,2,3,4,5,6,7,8 or 9    | 100.00 euros | <input type="radio"/> | 1,2,3,4,5,6,7,8 or 9    | 192.50 euros | <input type="radio"/> |
|          | 10                      | 80.00 euros  |                       | 10                      | 5.00 euros   |                       |
| 10       | 1,2,3,4,5,6,7,8,9 or 10 | 100.00 euros | <input type="radio"/> | 1,2,3,4,5,6,7,8,9 or 10 | 192.50 euros | <input type="radio"/> |
|          | ---                     | 80.00 euros  |                       | ---                     | 5.00 euros   |                       |



# INTERNET PAYMENT

Figure 1.5: Screenshot of the Risk Aversion Elicitation Page

WORLD CUP 2006  
TORLABOR SOCCER TRADING MARKETS

Please answer the questions below. Questions marked with a (\*) are mandatory. In case you feel uncomfortable with answering some of the non-mandatory questions, you can skip them. The [Cologne Laboratory for Economic Research](#) and the [Max Planck Institute of Economics](#) in Jena would be grateful if you would answer all questions honestly. Your assistance is integral to the success of the research conducted by the Cologne Laboratory for Economic Research and the Max Planck Institute of Economics in Jena.

Our [terms and conditions](#) apply, which implies that your data will be anonymized before the analysis.

Year of birth:

\* Your home country:

\* Your gender:  female  
 male

Your marital status:  married  
 single  
 divorced  
 widowed  
 other

\* Your degree of expertise in soccer is: low  high

\* Your degree of experience in trading markets is: low  high

How have you heard about the TorLabor soccer trading markets?:  invitation mail from an experimental lab  
 friends  
 newspaper  
 coincidental  
 information provided by another website  
 other

---

Who in your household would you consider to be primarily in charge of expenses and budget decisions?:  self  
 spouse  
 parent  
 other (specify)   
 do not know

How would you best describe your current employment situation?:  full-time employed (outside of school)  
 part-time employment (outside of school)  
 self-employment (outside of school)  
 unemployment  
 student only  
 work at school (research assistant, professor)  
 other

---

If you are a **student**, please answer the following questions

What describes your current student situation best?  full-time student  
 part-time student taking less than 12 hours per semester

Your field of studies:   
other

Your current semester:   
 undergraduate level  
 graduate level

Who is primarily responsible for your tuition and living expenses while you are attending university?  self  
 parent  
 shared between self and parent  
 scholarship / grant  
 loans  
 combination / other  
 not applicable

### 1.7.2 Ordered Probit Regression with Number of Safe Choices

We now take a look at the results of the ordered probit analysis of the number of safe choices. Note, that in this analysis we implicitly sort the data like HL. All  $A$  choices are assumed to be before all  $B$  choices. The graphs are slightly different from the graphs of the switching point analysis. Figure 1.6 (upper panel) shows the results of all data. Lab and Internet are not virtually identical any more but still not significantly different in the regression results. Table 1.7 shows the regression results. All main results are essentially identical for the analysis of the whole dataset. The most important result is even strengthened as the lab high stakes data already shows significantly more risk aversion in the entire dataset.

Contrary to the MSP analysis, there are hardly any differences between the entire and the consistent dataset, neither in the graph nor in the regression table. This is straightforward – the impact of inconsistently completed HL tables is minimized by the ordering of  $A$  and  $B$  choices already.<sup>37</sup> So we do not expect analyzes that compare an entire dataset to a consistent subset by the method of counting safe choices to yield any substantial differences. Maier and Ruger (2011) do this in a side note of their analysis and in line with our analysis do not find significant differences.

For the analysis of the consistent data, all results are actually identical between MSP and number of safe choices, as there is no inconsistent data to be treated differently. We can see this by comparing the respective Model 2 of Table 1.5 and Table 1.7. Graphically, we can see this from the respective lower panel of Figure 1.2 and Figure 1.6 which are identical.

---

<sup>37</sup>Given the regular pattern of HL table choices, a  $B$  choice has the biggest impact in one of the lower number rows. However, it is forced to come up in the higher number rows by the sorting method.

INTERNET PAYMENT

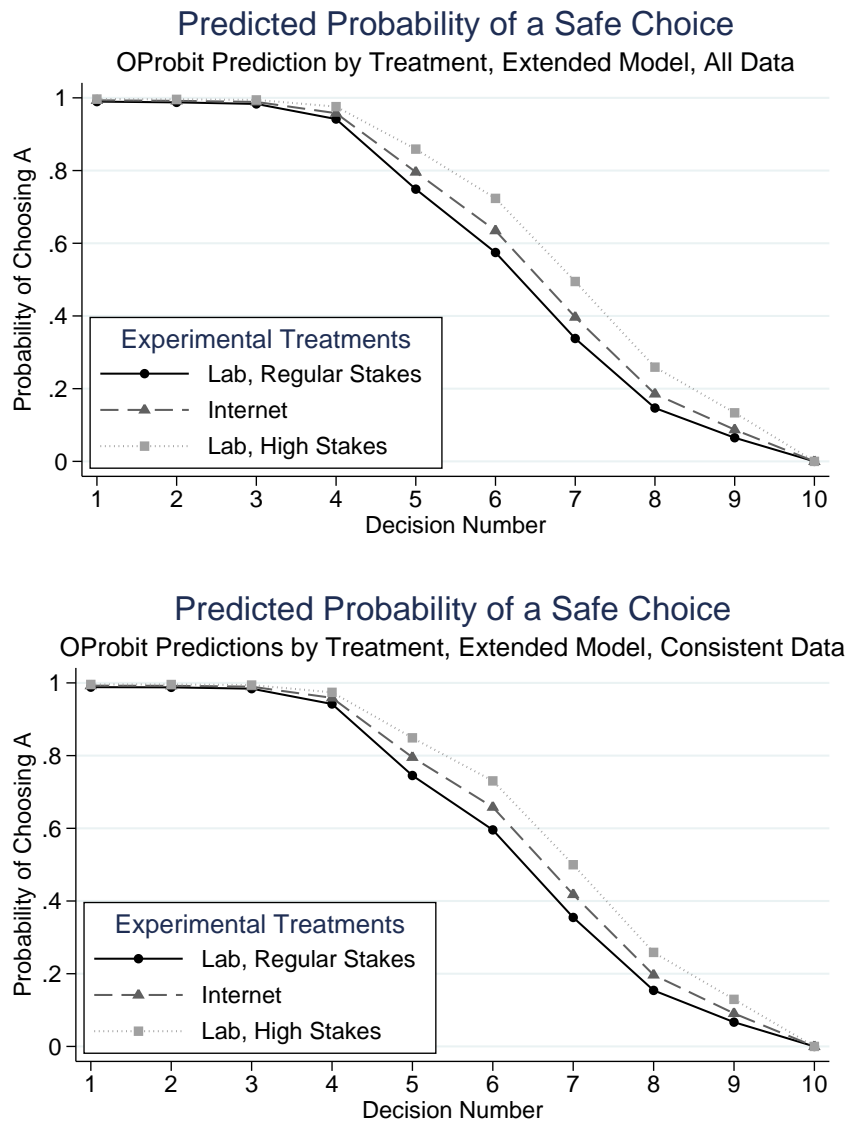
Table 1.7: Ordered probit estimates (Number of safe choices) of likelihood to choose lottery A - models with full demographics

|                           | Model 1     |           | Model 2         |           |
|---------------------------|-------------|-----------|-----------------|-----------|
|                           | All Data    |           | Consistent Data |           |
|                           | Coefficient | Std. Err. | Coefficient     | Std. Err. |
| Internet                  | 0.1571      | (0.1406)  | 0.1668          | (0.1435)  |
| Lab High Stakes           | 0.4073**    | (0.1965)  | 0.3746**        | (0.2016)  |
| Female                    | 0.0294      | (0.0422)  | 0.0501          | (0.0447)  |
| Age (at begin of studies) | -0.0069     | (0.0081)  | -0.0035         | (0.0089)  |
| Semester                  | 0.0036      | (0.0054)  | 0.0027          | (0.0056)  |
| <i>Field of studies</i>   |             |           |                 |           |
| Economics                 | 0.0840      | (0.0602)  | 0.0749          | (0.0619)  |
| Engineering               | 0.1290      | (0.0812)  | 0.1215          | (0.0848)  |
| Natural Science           | 0.3464***   | (0.0866)  | 0.3443***       | (0.0902)  |
| Law                       | 0.1198      | (0.0887)  | 0.2084**        | (0.0959)  |
| Other                     | 0.1278**    | (0.0518)  | 0.1832***       | (0.0544)  |
| <i>Home country</i>       |             |           |                 |           |
| China                     | 0.0310      | (0.2072)  | 0.0609          | (0.2213)  |
| Turkey                    | -0.1021     | (0.2155)  | 0.0977          | (0.2433)  |
| Russia                    | 0.2392      | (0.2228)  | 0.1964          | (0.2390)  |
| Poland                    | 0.1816      | (0.2309)  | 0.1716          | (0.2438)  |
| Bulgaria                  | 0.6493      | (0.3525)  | 0.4620          | (0.3687)  |
| Ukraine                   | 0.3170      | (0.2874)  | 0.4918          | (0.3297)  |
| Kazakhstan                | 0.1621      | (0.3239)  | 0.1205          | (0.3242)  |
| Other                     | 0.0166      | (0.0894)  | 0.1130          | (0.1008)  |
| <i>Charge of expenses</i> |             |           |                 |           |
| Parents                   | 0.0641      | (0.0604)  | 0.0685          | (0.0639)  |
| Spouse                    | -0.0697     | (0.2066)  | -0.0117         | (0.2281)  |
| Other                     | 0.1129      | (0.0982)  | 0.1420          | (0.1050)  |
| <i>Marital status</i>     |             |           |                 |           |
| Married                   | -0.1072     | (0.1473)  | -0.0671         | (0.1596)  |
| Other                     | -0.0309     | (0.1161)  | 0.0214          | (0.1281)  |
| cut1                      | -2.3244     | (0.2043)  | -2.1774         | (0.2649)  |
| cut2                      | -2.2520     | (0.2016)  | -2.1579         | (0.2643)  |
| cut3                      | -2.1363     | (0.1980)  | -2.0567         | (0.2615)  |
| cut4                      | -1.5745     | (0.1899)  | -1.4752         | (0.2543)  |
| cut5                      | -0.6709     | (0.1869)  | -0.5566         | (0.2520)  |
| cut6                      | -0.1854     | (0.1866)  | -0.1360         | (0.2517)  |
| cut7                      | 0.4242      | (0.1865)  | 0.4830          | (0.2518)  |
| cut8                      | 1.0608      | (0.1871)  | 1.1346          | (0.2525)  |
| cut9                      | 1.5275      | (0.1882)  | 1.6198          | (0.2536)  |
| Log likelihood            | -5,043      |           | -4,561          |           |
| Number of obs             | 2,709       |           | 2,460           |           |
| LR $\chi^2(23)$           | 35.53       |           | 37.82           |           |
| Prob > $\chi^2$           | 0.0460      |           | 0.0267          |           |
| Pseudo $R^2$              | 0.0035      |           | 0.0041          |           |

Notes: \*, \*\*, and \*\*\* indicate significance at the 10%, 5% and 1% level, respectively. The baseline category is “Laboratory”. Omitted variables are the ones with the highest frequency in their respective class: Business Administration (field of studies), Germany (home country), Self (charge of expenses), Single (marital status). In Model 2, only data from consistent choice profiles are included.

# INTERNET PAYMENT

Figure 1.6: Predicted safe choices



# Chapter 2

## What Makes them Tick

## Beliefs, Risk Aversion, and Final Prices in Prediction Markets<sup>0</sup>

### 2.1 Introduction

Explicit prediction markets exist for about 25 years and have steadily gained attention within the public and the economics profession. Through their final price they provide sound advice for the forecast of the respective future event. They are also usually better than most other forecasting tools. Up to now, the majority of the literature focuses on the application and the performance of such markets. Our focus is on the research regarding the price formation. This strand of literature identifies traders' beliefs and their risk aversion as the two most important driving factors of a prediction market's final price. But this literature relies mainly on theoretical analyses or simulated data for belief distributions and assumed utility functions for traders. Empirical data is only marginally used.

We add data generated in actual prediction markets – not through simulations – to this

---

<sup>0</sup>This chapter is based on joint work with Ben Greiner and Anthony Ziegelmeyer.

strand of literature. Our markets were set up during the 2006 FIFA Soccer World Cup. The uncertain events to be predicted were the outcome of each of the matches. Each match had one independent market. We introduce the distinction between the (theoretical) equilibrium price and the (empirical) final price of a market. With our dataset of risk aversion, beliefs, and final prediction market prices we can compare the mean belief, the equilibrium price, and the final price on a market.

Our results support the prevalent notion of no relevant differences between mean belief and equilibrium price. Furthermore, our final prices exhibit a considerable difference to the equilibrium prices. Additionally, in our markets, the final prices are the worst predictive statistics and are outperformed by the mean belief. Our contribution to the literature is the distinction between final and equilibrium price as well as the comparison of the predictive power of mean belief, equilibrium price, and final price.

Prediction markets are based in the general belief of economists that information is incorporated into markets through the price mechanism. This way, markets can aggregate (private) information that is dispersed across market participants. Hayek (1945) was the first one to write about this feature of markets. He describes the market as an information gathering system which displays *relevant* information to the market participants. The author gives the example of an increased tin market price. A trader does not need to know whether supply decreased or demand increased. The only relevant information to him is that the scarcity of tin has changed and this in turn increased the price in order to bring market supply and demand back to equilibrium.

But the price mechanism is not only useful for gathering information on current states of the world. It can also gather information on future states of the world as long as there are information and/or expectations available at present. Generally spoken, the price of any security in which contracting, payment, and delivery of the underlying do not take place at the same point in time (partly) reveals the contracting parties' information and beliefs about the future. With a market of fair thickness this information about the future should be incorporated into the current market price through trading. An example is given by

Roll (1984): He shows that prices of orange juice futures could not be predicted with the aid of weather forecasts (once it was controlled for institutional limits of price movement). An overview of experimental work considering the “Hayek Hypothesis” is given in Smith (1982).

Information aggregation in prices through trades is a prerequisite for the efficient market hypothesis (EMH). According to the EMH, an informationally efficient market is a market that reflects all information at the respective point in time through its prices. Formal foundations of this hypothesis emerged in Samuelson (1973) and Mandelbrot (1966). Among others, Fama (1965a), Fama (1965b), and Fama et al. (1969) further developed the EMH, until it was comprehensively presented in Fama (1970). Such a strong statement like the EMH sparked an extensive discussion on the efficiency of markets and the general possibility of efficient markets, which included Grossman (1976), Jensen (1978), Tirole (1982), and the seminal paper of Grossman and Stiglitz (1980). Literature reviews are Farmer and Lo (1999), and Malkiel (2003). Nowadays, the EMH is widely accepted as a “benchmark” (Fama, 1991) for the analysis of price movements on financial markets.

The EMH was incorporated into the development of markets that had the sole purpose of generating a forecast about future events. The first paper on this subject was Forsythe et al. (1992). In 1988, a prediction market<sup>1</sup> was implemented at the University of Iowa in order to predict the winner of the upcoming presidential election of George Bush (Senior) and Michael Dukakis.<sup>2</sup> Currently, the name of this platform is “Iowa Electronic Markets” (IEM)<sup>3</sup> and it is the oldest academical prediction market institution. The forecast gained from this market was strikingly accurate and outperformed opinion polls as the alternative forecasting instrument. Its essential design set a standard and is nowadays employed in all other prediction markets. Each potential outcome of a future event is assigned one respective contract that pays money related to the realization of the underlying outcome.

---

<sup>1</sup>Alternative names include “news futures”, “information markets”, and “forecasting markets”.

<sup>2</sup>To be semantically correct, these markets deliver “forecasts”, not “predictions”. “Prediction” implies the eventual realization of the predicted outcome. This is not always true for prediction markets. Like in the literature of prediction markets we will use these terms interchangeably.

<sup>3</sup><http://tippie.uiowa.edu/iem/index.cfm>

These contracts can be freely traded on a double auction platform. Resulting prices deliver information of the estimated probability of occurrence. Although the average trader in Forsythe et al. (1992) exhibited (judgement) biases, the market worked well. The authors identify the “marginal traders” as the reason for the positive market performance. These traders were defined as exhibiting overproportionally frequent final price setting behavior. The average marginal trader invested more money than the average non-marginal trader, had a higher return on his invested money, showed a more successful trading strategy conditional on news quality, and was less prone to biases.

From that point of time, prediction markets have been frequently implemented not only to predict election outcomes. Commercial prediction markets include the Hollywood Stock Exchange<sup>4</sup>, which focuses on movie related underlyings and the commercial leader Intrade<sup>5</sup>, which covers a broad variety of events like elections, gasoline prices, and the discovery of a supersymmetric particle.<sup>6</sup> Prediction markets are still regularly set up for academic research. The IEM are the most frequent academic markets. Even companies like HP, Siemens, and Pfizer began to use prediction markets in order to aggregate information to improve decision making processes. Usually, these markets forecast sales figures like in Chen and Plott (2002).

The majority of research on prediction markets is on the astonishing accuracy, supporting the thesis that such a prediction market price usually constitutes the best available forecasting instrument. As in Berg et al. (1996) and Oliven and Rietz (2004), the research focus of this strand of literature is on generating knowledge on design elements which enhance prediction markets’ forecasts. For example, this research tries to identify demographic characteristics of the marginal trader. Knowing more about him seems to be key to understand the mechanism through which prediction markets perform so well. The focus of the second strand of literature is different. It is on characteristics of the

---

<sup>4</sup><http://www.hsx.com/>

<sup>5</sup><http://www.intrade.com>

<sup>6</sup>Essentially, all betting markets are implicitly also prediction markets. They provide similar incentives. Therefore one can derive predictions from betting quotas.



entire set of market participants, with beliefs and risk aversion as the two most important dimensions.

Manski (2006) shows that for risk-neutral traders with heterogeneous beliefs the final price gives merely partial information on the central tendency of beliefs. His results triggered further theoretical and data-based literature. Adams (2006) adds a different kind of learning than Manski to Manski's model and concludes that this makes the final price converge to the mean of beliefs. Gjerstad (2005) theoretically demonstrates that traders' risk aversion and their beliefs can be important drivers of the final price. In his theoretical analysis he concludes that for empirically validated degrees of relative risk aversion and "plausible" belief distributions, the equilibrium price is "very near the traders' mean belief". This is corroborated by Wolfers and Zitzewitz (2007). They develop a parsimonious model and test it with simulated data for utility functions and belief distributions as well as with field data from prediction markets and sports betting markets. They also find that "prediction markets prices typically provide useful (albeit sometimes biased) estimates of average beliefs about the probability an event occurs". Similarly, in their analysis of simulated belief data Fountain and Harrison (2011) find equilibrium prices in markets with traders with log or CRRA utility functions to be rather close to the mean belief as long as the belief distribution is unimodal and trader characteristics like wealth and time preferences are orthogonal to beliefs. The focus of Ottaviani and Sørensen (2012) is rather on testing financial asset pricing theories in a dynamic setting, but their model can be interpreted as the most general prediction market model so far. It also includes information gathering of the traders through other sources than the market. Contrary to the other papers, they regard a marginal trader. They find under-reaction of the equilibrium price to information if traders have an investment limit or DARA risk preferences.

Schmidt and Werwatz (2002) set up prediction markets in the context of the UEFA European Soccer Championship 2000. These markets outperformed quotas from an online betting platform and a random predictor. Unpublished work on soccer prediction markets

include markets at CREED of the University of Amsterdam for the 1994 FIFA Soccer World Cup and markets at the Humboldt University of Berlin for the 1998 FIFA Soccer World Cup.

Finding out how prices on prediction markets are generated not only helps us to discover how these markets work. Through the similarity of prediction markets to regular double auction markets, analyzing these markets also aids our understanding of market price formation in general. Prediction markets have the advantage of richer datasets on traders and prices as usual financial markets. They also reveal the fundamental value of a contract at a certain point of time in contrast to many assets on regular financial markets.

Our experiment is the first study to analyze the interplay of risk aversion and beliefs with the aid of data elicited from human traders. We set up prediction markets which give us final prices to analyze. Additionally, we elicited the participants' risk aversion and beliefs. All other studies of the related strand of literature either relied on theoretical analyses or used (partly) artificially created datasets. Data created in actual prediction markets was never used in more but a supplementary way in this strand of literature. Furthermore, no study so far analyzes a dataset as exhaustive as ours.

## **2.2 Research Design**

### **2.2.1 General Setup**

Our research question required the elicitation of the participants' risk aversion, their beliefs about the underlying events and the results of prediction markets. Details of the general setup and the risk elicitation procedure are described in the first chapter of this thesis. We will focus on the main design elements as well as the belief elicitation procedure and the prediction markets themselves.

Following our research question, there were three incentivized tasks for the participants.

## WHAT MAKES THEM TICK

The first one was the HL table, the second one was the submission of probabilistic beliefs of match outcomes and the third one was the trading on prediction markets. Five participants were paid according to their choices in the risk elicitation task, another five were paid according to their beliefs, and another 20 were paid according to their performance in one prediction market. A participant who was selected for payment would only be paid for one of the three tasks. To suppress hedging motives, this was clearly stated in the instructions.

The FIFA Soccer World Cup 2006 was a big event which generated dozens of smaller events with uncertain outcomes – the single matches. We recruited 3,582 participants which were allocated across 16 matching groups. A registering participant was allocated to the matching group with the least number of participants and randomly if there was more than one matching group. This way we achieved similar participant numbers across all matching groups. Eight matching groups had classical double-auction prediction markets as the trading platform. Participants of the other eight matching groups traded on pari-mutuel betting markets. This analysis focuses on the double auction markets to which 1,790 participants were allocated.

64 matches were played in this world cup. We set up a market for each of the matches and each of the matching groups, which gives us 512 markets overall. The first 48 matches were played in a group stage (first round) in which each match of Team A vs. Team B had three possible outcomes: Team A wins, Team B wins, tie. Team A can also be called “Team of record”. The 16 matches of the following playoff stage (second round) did not include “tie” as a possible outcome. Matches of the second round were not ended before a winner was determined.<sup>7</sup> In this analysis we only regard games of the second stage as the entire prediction market literature focuses on markets with two outcomes. Analyzing

---

<sup>7</sup>A game lasted two times 45 minutes, plus extra time of two times 15 minutes if no winner was determined after 90 minutes, plus a penalty shootout if no winner was determined after extra time. Extra time and penalty shootout were only used in the second stage. The exact rules of the FIFA tournament can be found under this stable URL:

<http://eur.i1.yimg.com/eur.yimg.com/i/eu/fifa/regen.pdf>

More information can be gathered on the FIFA homepage:

<http://www.fifa.com>

markets with more than two assets will be left for future research.

We chose the world cup as the uncertainty generating event, because of a high soccer interest in Germany. This interest was even strengthened as Germany was the host country of the world cup. Many people would have different opinions about the matches, which motivates participation in the experiment and trading on the respective markets. As stressed by Wolfers and Zitzewitz (2004) and Snowberg et al. (2012) this is an important prerequisite to generate “thick” markets and actually incorporate information into the market prices.

Any person could register on the homepage of the experiment.<sup>8</sup> During the registration process, participants submitted their name, e-mail address, a username and a password. Afterwards we elicited their risk aversion through a multiple price list (MPL) designed by Holt and Laury (2002) (HL) and described in detail in the first chapter of this thesis. The registration procedure was completed by a questionnaire for demographical data. Mandatory information were sex and home country. All participants selected for payment were informed via e-mail and paid by wire transfer after the World Cup final.

After the registration process the participants were free to enter any market. Before a participant could enter a market we elicited his beliefs about the probabilities of the two potential match outcomes via a linear scoring rule. Past submission he entered the market. In order to increase the number of entered markets we implemented an “eight markets rule”. If one participant was randomly determined for payment according to his market performance, the number of his entered markets was checked. Had he entered at least eight markets, one of these would be randomly chosen for payment. Had he entered less than eight markets, dummy markets with a performance of 0 € to sum up to eight markets were employed.

---

<sup>8</sup>[www.torlabor.de](http://www.torlabor.de)

## 2.2.2 Markets

Our market design closely follows the established design of the IEM. A market for a respective match usually opened 74 hours and ended 2 hours before the scheduled starting time of the respective match. So the usual trading time was 72 hours. Because of the tight schedule of the world cup, the teams of three second round matches were not determined early enough for such a runtime. These markets ran for 66 hours.

On every market, each trader received an endowment of 100 € but no contracts. Each possible outcome of the match was assigned an Arrow-Debreu-security.<sup>9</sup> The underlying events of these securities were mutually exclusive and collectively exhaustive. A security would pay 1 € if the event occurred and 0 € otherwise. At any point during the runtime of the respective market, traders could buy and sell “bundles” from and to the bank. A bundle contained one of each kind of securities, had the constant price of 1 € throughout the runtime of the market and could only be traded with the bank. Single securities could only be traded among the market participants. Neither money nor shares were transferable between markets for a certain trader. It was also impossible to transfer money or shares directly from one trader to another. The securities could be freely traded on a double auction platform in price increments of 0.01 €. A negative account balance, short selling, and creating new securities were impossible. Trading was not mandatory. Participants could enter the market and do nothing. This would ensure them with a payoff of 100 € in case they were drawn for payment according to their performance of this market.

The main difference of TorLabor to the IEM real money markets is that our participants did not have to invest their own money but received an endowment for each entered market. Many countries do not allow real-money prediction markets or have ambiguous or unfavorable legal settings. Even if real-money prediction markets are in accordance to the laws of the respective jurisdiction, often special permissions are advisable or nec-

---

<sup>9</sup>In finance, these securities are named “binary calls”. This is also how they are legally defined in most cases. As such, they fall into the legal category of “options”. We called them “shares” throughout the experiment in order to avoid confusion of the participants.

essary for such markets. Regularly this also applies for zero-sum pure research markets at universities. In fact, just recently (in November 2012) the US Commodity Futures Trading Commission (CFTC) charged Dublin-based Intrade “with violating the CFTC’s off-exchange options trading ban and filing false forms with the CFTC”. This effectively bans USA residents from the platform.<sup>10</sup> Arrow et al. (2008) propose reforms to simplify the legal process of a prediction market setup.

To minimize legal issues, we refrained from setting up a real-money market. When a participant was determined for payment according to a certain market, he was paid his winnings on this respective market in actual money (exchange rate 1:1). Not using real money of the traders themselves marks a departure from the usual incentivized prediction market design. In case of influence on participant behavior this could confound our results. However, Servan-Schreiber et al. (2004) and Slamka et al. (2008) did not find significantly different performance of real-money and play-money markets. So we assume that our markets perform similarly to real-money markets.

The equal endowment for each trader on each market not only had the practical benefit of prevented jurisdictional problems. This design element also implemented an assumption of the relevant prediction market literature – orthogonality of beliefs and trader wealth at the beginning of a market. Traders were free to invest as much of their endowment as they wished. But their wealth at the beginning of each market was set to 100 €. Traders could increase their wealth through successful trading during the market runtime and therefore invest more than 100 € eventually. But they started out as being equally rich.

---

<sup>10</sup>The CFTC’s press release can be found here:  
<http://www.cftc.gov/PressRoom/PressReleases/pr6423-12>  
Intrade’s statement can be found here:  
<http://www.intrade.com/news/id/782>  
Intrade is currently searching for a way to solve the legal issues in order to continue to offer their service to USA residents.

Table 2.1: The ten paired lottery-choice decisions

| Row | Lottery $A$              | Lottery $B$                | $E_A - E_B$ | CRRA coefficient if row was last A choice, below all B choices |
|-----|--------------------------|----------------------------|-------------|--|
| 1   | { 100€, 0.1 ; 80€, 0.9 } | { 192.50€, 0.1 ; 5€, 0.9 } | 58.25€      | [-1.71; -0.95]   |
| 2   | { 100€, 0.2 ; 80€, 0.8 } | { 192.50€, 0.2 ; 5€, 0.8 } | 41.50€      | [-0.95; -0.49]   |
| 3   | { 100€, 0.3 ; 80€, 0.7 } | { 192.50€, 0.3 ; 5€, 0.7 } | 24.75€      | [-0.49; -0.14]   |
| 4   | { 100€, 0.4 ; 80€, 0.6 } | { 192.50€, 0.4 ; 5€, 0.6 } | 8.00€       | [-0.14; 0.15]  |
| 5   | { 100€, 0.5 ; 80€, 0.5 } | { 192.50€, 0.5 ; 5€, 0.5 } | -8.75€      | [0.15; 0.41]   |
| 6   | { 100€, 0.6 ; 80€, 0.4 } | { 192.50€, 0.6 ; 5€, 0.4 } | -25.50€     | [0.41; 0.68]   |
| 7   | { 100€, 0.7 ; 80€, 0.3 } | { 192.50€, 0.7 ; 5€, 0.3 } | -42.25€     | [0.68; 0.97]   |
| 8   | { 100€, 0.8 ; 80€, 0.2 } | { 192.50€, 0.8 ; 5€, 0.2 } | -59.00€     | [0.97; 1.37]   |
| 9   | { 100€, 0.9 ; 80€, 0.1 } | { 192.50€, 0.9 ; 5€, 0.1 } | -75.75€     | [1.37; $\infty$ )  |
| 10  | { 100€, 1.0 ; 80€, 0.0 } | { 192.50€, 1.0 ; 5€, 0.0 } | -92.50€     | non-monotone   |

Notes:  $E_A - E_B$  denotes the expected payoff difference between lottery A and lottery B. Participants were only shown the information of the first three columns. They had no explicit information on  $E$  or the CRRA coefficient.

### 2.2.3 Risk Elicitation

The risk aversion elicitation procedure closely follows Holt and Laury (2002). We employed their “very high” payment scheme whilst replacing “\$” with “euros”. Table 2.1 illustrates the payoff matrix presented to our subjects.<sup>11</sup> This way, the potential payoffs ranged from 5.00€ to 192.50€ with intermediate payoffs at 80€ and 100€. Choosing this range was guided by the aim to align the payoff range and the expected payoff of random (respectively no) actions in all three tasks of the experiment. In case the absolute values of the lotteries mattered for the elicitation of the subjects’ risk aversion, we minimized confounds in our estimation with this strategy.

In each of the ten rows, a participant had to choose between lottery A and lottery B. A rational risk neutral individual should pick lottery A in the first four rows and pick lottery B in each consecutive row from row five on. Switching earlier is related to risk-loving behavior, switching later is related to risk-averse behavior. Switching from lottery B to lottery A in a subsequent row as well as choosing lottery A in row 10 is called “inconsistent” behavior. For an extensive description and discussion of the risk elicitation procedure as well as the estimation of individual parameters of risk aversion, see the first

<sup>11</sup>At the time of the experiment, 1€ was worth about US\$1.25.

chapter of this thesis.

With the help of the results of this task we will estimate the risk aversion parameter of the utility function. For the analysis we assume a power utility function with constant relative risk aversion (CRRA). The utility function we use throughout the analysis is characterized by:

$$U(x) = \begin{cases} \frac{x^{(1-r)}}{1-r} & \text{for } r \neq 1 \\ \ln x & \text{for } r = 1 \end{cases} \quad (2.1)$$

The parameter to be estimated for each individual is  $r$ . An  $r > 0$  represents risk aversion, an  $r = 0$  risk neutrality, and an  $r < 0$  risk love. In the data analysis we do not have to consider cases of log utility as no participant has  $r = 1$  in our estimations. The estimations are conducted via interval regression and are described in the first chapter of this thesis.

## 2.2.4 Beliefs

Before entering a market, all participants had to report their beliefs of the likelihood (their subjective probability)<sup>12</sup> of each of the two teams winning the respective match. The literature distinguishes between prior and posterior beliefs, which differ in the point of time a trader holds them - before he enters the market or after the trading period of the market has ended. According to this taxonomy, we elicited prior beliefs. Most theoretical papers only regard posterior beliefs in their analysis. In this respect it would have been preferable to elicit posterior beliefs. But for the market analysis we need belief data for at least the vast majority of individuals who entered the respective market. Without their beliefs, we would lose the entire market as an observation.

Incentives were given through a linear scoring rule. After a participant was randomly chosen for payment according to his beliefs, one of his entered markets was randomly chosen. The participant would be paid the percentage he had assigned to the eventual outcome of 100 €. On the one hand, such a linear scoring rule will let risk aversion play

---

<sup>12</sup>“Subjective probability” and “belief” will be used interchangeably throughout.



a role in the submitted beliefs so that subjects usually do not report their true beliefs. On the other hand, this rule is easy to understand. Alternatively, we could have used a “proper” scoring rule. Such a rule reaches its maximum value at the true beliefs of participants. It is “strictly proper” if this is the unique maximum. Quadratic scoring rules are the most popular strictly proper scoring rules.<sup>13</sup>

However, proper scoring rules are not easily comprehensible for regular participants. Nelson and Bessler (1989) test this for a quadratic scoring rule in a set of risk neutral participants. According to them, the rule requires practice rounds by the participants to elicit true beliefs. More importantly, by construction it requires risk neutrality of the participants. So even beliefs elicited by proper scoring rules usually have to be corrected for risk aversion. Offerman et al. (2009) provide an extensive analysis and discussion of this topic. Nelson and Bessler (1989) show that the forecaster’s subjective probabilities enter the maximization problem linearly and therefore lead to corner solutions for risk-neutral participants. All maximum beliefs are reported as any value of the interval  $[0, 1]$  (while sufficing non-negativity for each belief and a unity sum) and all non-maximum beliefs are reported as 0.

Andersen et al. (2010) discuss quadratic and linear scoring rules and compare probabilities elicited by such rules. They do not find differences between the belief distributions after correcting for risk.<sup>14</sup> We did not expect our participants to be risk neutral and wanted to maximize comprehension of the rule in order to foster participation. Therefore we chose the linear scoring rule.

Formally, a linear scoring rule can be described in the following way. The true state of nature has to be forecasted. There can be  $S$  different states of nature.  $s$  denotes a single state of nature, so  $s \in (1, 2, \dots, S)$ . All states of nature are mutually exclusive and collectively exhaustive.  $q_s$  denotes the forecasters stated probability for state  $s$ . This is the figure that the researcher elicits from the subjects. We assume  $p_s, q_s \geq 0 \forall s, \sum_{s=1}^S p_s = 1$ ,

---

<sup>13</sup>Other examples include logarithmic and spherical scoring rules.

<sup>14</sup>It may still be the case that both scoring rules elicited the same “untrue” belief of a participant.

and  $\sum_{s=1}^S q_s = 1$ . A forecaster has the subjective probability vector  $(p_1, p_2, \dots, p_S)$ . His stated probability vector is denoted by  $(q_1, q_2, \dots, q_S)$ . The payoff  $L$  (100 € in our setting) is multiplied by the stated probability of the eventual outcome (the true state). The forecaster's final payoff (FP) is computed by:

$$FP = \sum_{s=1}^S \mathbf{1}_A(s) L q_s \in [0, L] \quad (2.2)$$

Where “ $\mathbf{1}_A(\text{condition})$ ” denotes the indicator function. It has value 1 if the stated condition is true and 0 otherwise. The forecaster reaches the maximum payoff if he assigns probability 1 to the eventual outcome and the minimum payoff if he assigns probability 0 to the eventual outcome.

The linear scoring rule does not elicit true beliefs of the participants for any utility function other than log utility. Therefore, elicited beliefs of individuals without log utility must be corrected for risk aversion in order to use them for further analyses with participants' true beliefs. In the following, we describe the participants' maximization problem and derive functions to obtain subjective probabilities from stated probabilities.

The forecaster, for whom we assume probabilistic sophistication in the sense of Machina and Schmeidler (1992), has to maximize his final payoff over his stated probabilities  $q_s$ :

$$\max_{q_1, q_2, \dots, q_S} \sum_{s=1}^S p_s U(Lq_s) \text{ s.t. } q_i \geq 0 \forall s = 1, 2, \dots, S \text{ and } \sum_{s=1}^S q_s = 1 \quad (2.3)$$

As long as not stated otherwise, we follow the majority of the theoretical prediction market literature we discuss and will assume  $S = 2$  in the further analysis. So we have:

$$\max_{q_1, q_2} \sum_{s=1}^2 p_s U(Lq_s) \text{ s.t. } q_i \geq 0 \forall s = 1, 2 \text{ and } \sum_{s=1}^2 q_s = 1 \quad (2.4)$$

This leads to:

$$\begin{aligned}
 p_1 &= \frac{U'(Lq_2)}{U'(Lq_1) + U'(Lq_2)} \\
 &= \frac{U'(L(1 - q_1))}{U'(Lq_1) + U'(L(1 - q_1))} \\
 p_2 &= 1 - p_1
 \end{aligned}$$

With our utility function (and  $r \neq 0$ ) this can be reformulated to:

$$\begin{aligned}
 p_1 &= \frac{(1 - q_1)^{(-r)}}{q_1^{(-r)} + (1 - q_1)^{(-r)}} \\
 p_2 &= 1 - p_1
 \end{aligned} \tag{2.5}$$

We assume  $p_1 = q_1$  if  $q_1 \in \{0, 1\}$  for  $r > 0$  and  $p_1 = 1 - q_1$  if  $q_1 \in \{0, 1\}$  for  $r < 0$ . Here we can see that log utility ( $r = 1$ ) yields a truthful revelation of subjective probabilities.

According to Nelson and Bessler (1989), for  $p_1 \neq 0.5$  we only receive extreme values (0 or 1) of  $q_1$  for risk neutral subjects ( $r = 0$ ). Additionally, for  $p_1 = 0.5$  the forecaster submits  $q_1 \in [0, 1]$ . A risk neutral individual is indifferent between any two vectors of stated probabilities if his  $p_1 = 1/2$ .<sup>15</sup> So instead of Function 2.5 the following applies:

$$\begin{aligned}
 p_1 &\in [0, 1] \text{ for } q_1 \in [0, 1] \\
 p_2 &= 1 - p_1
 \end{aligned}$$

Therefore we cannot learn anything from the stated probability of a participant.

However, if we assume that  $p_1 = 0.5$  iff  $q_1 = 0.5$ . We can derive more information from

---

<sup>15</sup>The general result is verbally presented above.

the stated probability:

$$p_1 \in \begin{cases} [0, 0.5) & \text{for } q_1 = 0 \\ [0.5, 0.5] & \text{for } q_1 = 0.5 \\ (0.5, 1] & \text{for } q_1 = 1 \end{cases} \quad (2.6)$$

$$p_2 = 1 - p_1$$

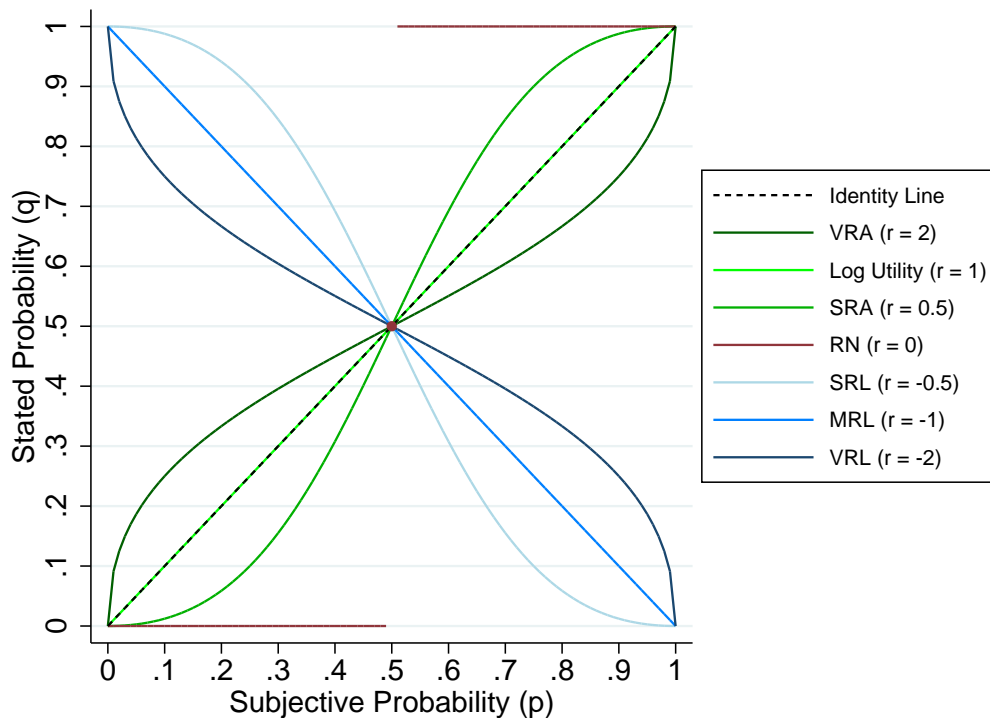
Still, from  $q_1$  we cannot derive more information on the individual's belief but the relation of  $p_1$  to 0.5.

The inverse functions of Function 2.5 and Function 2.6 are graphically illustrated in Figure 2.1. This figure follows the usual presentation form of the literature, e.g. Offerman et al. (2009) for a quadratic scoring rule. It is presented from the forecaster's perspective and displays  $q$  as a function of  $p$ , instead of  $p$  as a function of  $q$  like in the analysis above. It also refrains from subscript  $s$ . As long as not necessary for comprehension we will drop the subscript as well from here on.

As the graphs are point symmetric to  $(0.5, 0.5)$  we only discuss the cases in which  $p \geq 0.5$ . Results for  $p < 0.5$  can be derived accordingly. A point on the identity line means that stated probability and subjective probability coincide. This is true for all graphs at  $p = 0.5$ , by assumption for risk neutral subjects.

A very risk averse subject (defined by  $r > 1$  and depicted with  $r = 2$  in our graphical example) states probabilities biased towards 0.5. So  $q < p$  and  $q > 0.5$  for  $p > 0.5$ . His stated probabilities have to be corrected towards 1. As mentioned before, a subject with log utility will state true probabilities – his graph matches the identity line. A slightly risk averse subject (defined by  $0 < r < 1$  and depicted with  $r = 0.5$ ) states probabilities biased towards 1. So  $q > p$  and  $q > 0.5$  for  $p > 0.5$ . His stated probabilities have to be corrected towards 0.5. A risk neutral subject delivers probabilities of 1 for  $p > 0.5$ . Apart from  $p = 0.5$  (by assumption), it is impossible to derive his true probabilities with the aid of his utility function and his stated beliefs. For risk averse and risk neutral subjects the

Figure 2.1: stated probability vs. subjective probability in a linear scoring rule for different degrees of constant relative risk aversion



Notes: Each graph depicts the transformation of subjective into stated probability for one degree of constant relative risk aversion. VRA stands for very risk averse, Log Utility is the case for  $r=1$ , SRA stands for slightly risk averse, RN means risk neutral, SRL means slightly risk loving, MRL is for medium risk loving, and VRL for very risk loving.

relation of the probability towards 0.5 is always identical before and after correction.

The latter property is flipped for risk loving subjects ( $r < 0$ ). The relation of the probability towards 0.5 is after correction always the opposite of what it was before correction (apart from  $p = q = 0.5$ ). So a risk loving subject always states a probability  $q < 0.5$  for the event he actually rates as more likely ( $p > 0.5$ ) than the complementary event. Slightly risk loving subjects (defined by  $0 > r > -1$  and depicted with  $r = -0.5$ ) submit probabilities biased towards 0. Therefore,  $q < p$  and  $q < 1 - p$  and  $q < 0.5$  for  $p > 0.5$ . An individual with  $r = -1$  states the subjective probability of the complementary event.  $q = 1 - p$  and  $q < 0.5$  for  $p > 0.5$ . A very risk loving individual (defined by  $r < -1$  and depicted with  $r = -2$ ) states probabilities biased towards 0 as well.  $q < p$  and  $q > 1 - p$  and  $q < 0.5$  for  $p > 0.5$ . The stated beliefs of all risk loving participants have to be corrected towards 1.

The translation function graphs approach the horizontal line at  $q = 0.5$  when  $r$  approaches  $\infty$  or  $-\infty$ . The extreme points for  $p \in \{0, 1\}$  never change. So as an artefact, for extreme risk aversion and extreme risk love, the graphs approach each other.

## 2.3 Hypotheses

Following the research of Gjerstad (2005), Wolfers and Zitzewitz (2007), and Fountain and Harrison (2011) we set up the hypothesis that the final prices of the prediction markets are not substantially different from the mean belief of the subjects who entered the market.

### **Hypothesis 1 [MSB vs. MCB vs. EP]**

*There are no substantial differences between the mean corrected belief and the equilibrium price.*

Up to now, the literature regards the equilibrium price as the final price. It does not discuss the potential of differences between these prices but implicitly or explicitly assumes that these prices are identical. This leads to our second hypothesis.

### **Hypothesis 2 [MCB vs. EP vs. FP]**

*There are no differences between the equilibrium price and the final price.*

There is no established literature comparing the predictive accuracy of the mean stated belief, the mean corrected belief, the equilibrium price and the final price. We therefore hypothesize that there are no differences in the predictive accuracy.

### **Hypothesis 3 [Predictive Accuracy]**

*There are no differences in the predictive accuracy between the mean stated belief, the mean corrected belief, the equilibrium price, and the final price.*

## 2.4 Results

### 2.4.1 Descriptives

In contrast to the first chapter, in which we use the entire set of participants, we only use the internet participant for the following analysis. Here we focus on the analysis of prediction markets, and these were only set up in the Internet treatment. Table 2.2 shows the demographics of the Internet participants. This set of participants is used in Section 2.4.2. A random subset of 1,790 of these participants was allocated to the matching groups in the prediction market treatment. In the market analysis we only regard these participants.

Table 2.2: Demographic Characteristics of Participants

|                                   |             |
|-----------------------------------|-------------|
| Number of participants            | 3,582       |
| Avg. response rate                | 96.8%       |
| Age <i>Avg. (StdDev)</i>          | 25.6 (5.8)  |
| Sex <i>male</i>                   | 62.5%       |
| Marital Status                    |             |
| <i>Single</i>                     | 90.1%       |
| <i>Married</i>                    | 6.1%        |
| <i>Other</i>                      | 3.8%        |
| Charge of expenses                |             |
| <i>Self</i>                       | 81.9%       |
| <i>Parents</i>                    | 11.3%       |
| <i>Spouse</i>                     | 1.1%        |
| <i>Other</i>                      | 5.6%        |
| Most frequent countries of origin |             |
| <i>Germany</i>                    | 89.9%       |
| <i>China</i>                      | 1.0%        |
| <i>Turkey</i>                     | 0.8%        |
| <i>Russia</i>                     | 0.8%        |
| <i>Poland</i>                     | 0.6%        |
| <i>Other</i>                      | 7.1%        |
| Share of students                 | 79.8%       |
| For students:                     |             |
| Semester <i>Avg. (StdDev)</i>     | 6.00 (3.75) |
| Field of Studies                  |             |
| <i>Bus. Adm.</i>                  | 24.5%       |
| <i>Economics</i>                  | 18.8%       |
| <i>Engineering</i>                | 8.0%        |
| <i>Natural Science</i>            | 6.9%        |
| <i>Law</i>                        | 6.4%        |
| <i>Other</i>                      | 35.4%       |

Table 2.3 displays descriptive summary statistics on the prediction markets. We had 8 matching groups and 16 matches, which gave us 128 markets overall. The average number of entered traders is 86.2, so for each markets we do not obtain the beliefs from just a few traders. On average, traders submitted 175.57 offers for both assets, which resulted in 72.74 average trades on these markets. Although a few of the markets did not attract much interest of the traders, most markets had a fair level of activity. The figures on the bundles give us a proxy for the supplies of the assets. One bundle consisted of one of each assets. Traders could buy and sell bundles from and to the bank. For one given market, the current number of bundles started at 0 and usually increased with time as more traders entered and bought bundles. The average number of bundles averages the current number of bundles over all levels for one market. All figures had a negative time trend from one match to the next one.

Table 2.3: Summary Statistics for Markets

|                     | Mean     | Std. Dev. | Min.   | Max.    |
|---------------------|----------|-----------|--------|---------|
| Entered Traders     | 86.20    | 8.63      | 63     | 106     |
| Submitted Offers    | 175.57   | 49.02     | 77     | 327     |
| Trades              | 72.74    | 22.85     | 25     | 131     |
| Avg. No. of Bundles | 1,682.50 | 341.12    | 846.13 | 2598.32 |

## 2.4.2 Data Imputation

In our prediction market analysis we do not exclude anyone because of an inconsistency in the HL table or – like in the first chapter – incomplete demographic characteristics. Here we have to use the entire set of participants. The reason is the same as for the elicitation of prior instead of posterior beliefs. Markets in which we do not know the utility function and beliefs of all participants cannot be included in the analysis of the relation of beliefs and final prices.

To fill up the demographic data of participants who did not submit all of their characteristics we have to perform item imputation (Rubin (1987)). Specifically, we use “nearest neighbor hot deck imputation”. The nearest neighbor of an incomplete data point is de-



terminated via Euclidean distance of the items observed for both data points. Afterwards the missing item is donated from the matched data point to the incomplete data point. We search through the observations of our own dataset (hot deck) instead of comparing our data to another dataset (cold deck). Such an imputation mechanism systematically underestimates variability as it treats the missing items of the data points as if they were elements of the set of known item values (or convex combinations of these values) and known with certainty. As an upside, this method does not make any distributional assumptions and is relatively easy to implement. Furthermore, the imputed dataset can be used just like a non-imputed complete dataset.

The entire set of participants described in Table 2.2 was used in the data imputation. As participants were randomly allocated to one of the treatments (pari-mutuel or prediction market) we assume the demographics to be distributed similarly. However, we keep the imputation database as extensive as possible in order to optimize the imputation procedure.

### 2.4.3 Risk Aversion and Belief Correction

We estimate the participants' coefficients of risk aversion like in the first chapter of this thesis through an interval regression. We neither drop inconsistent participants nor do we – thanks to the imputation – need to drop participants with incomplete demographic information. Again, we have two models. The “short model” (SM) includes only sex and age as demographic variables, while the “extended model” (EM) includes all variables listed in Table 2.2.

Estimation results from empirical and experimental research like Binswanger (1980), Hansen and Singleton (1982), Goeree et al. (2003), or Bajari and Hortaçsu (2005) yield an  $r$  in the range of  $[-0.5, 1.8]$ . Goeree et al. (2002) provide an extensive overview on such studies and show that estimation results are mostly in the range of  $[0.3, 1]$  for individuals similar to our pool of participants. This means slight risk aversion for the entire

population of traders.

Table 2.4: Estimates of  $r$  for different models

|             | Mean | Std. Dev. | Min.  | Max. |
|-------------|------|-----------|-------|------|
| Intreg SM   | 0.64 | 0.05      | 0.32  | 0.72 |
| Intreg EM   | 0.64 | 0.10      | 0.24  | 1.06 |
| HL Midpoint | 0.52 | 0.62      | -1.71 | 1.37 |
| MLE SM      | 0.60 | 0.01      | 0.57  | 0.69 |
| MLE EM      | 0.61 | 0.05      | 0.45  | 0.78 |

The interval regression models deliver  $r$  estimation data shown in the upper panel of Table 2.4. The short model estimates the individual  $r$  in  $[0.32, 0.72]$ , while the extended model estimates  $r$  in  $[0.24, 1.06]$ . This is in line with our expectation based on previous studies. The estimated  $r$  is positive for each participant in both models – we only have risk averse participants in our dataset. No participant has an  $r = 1$ , so we do not have any participant with log utility. Apart from extreme values, we have to correct all beliefs of the participants. Only one single participant has an  $r > 1$  and is therefore very risk averse by our definition in Section 2.2.4. His beliefs are marginally biased towards 0.5 as his risk aversion coefficient is only marginally greater than 1. Therefore, the graph of the “probability translation function” of most participants should look similar to the bright green graph (SRA) in Figure 2.1. We expect our participants to state beliefs biased towards the extremes in all cases but one.

As robustness checks for the prediction market analysis, we used three other approaches to estimate the participants’ risk aversion. One very coarse but simple measurement of risk aversion in which participants were simply assigned an  $r$  which was the average of their minimal  $r$  and their maximal  $r$  determined by their lottery decisions in the HL table. Participants for which we had only one of these figures were assigned this figure. Participants for which we had neither the minimum nor the maximum were assigned an  $r$  of 0. This was the case for 32 participants. The two other approaches were maximum likelihood estimations (MLE) with Fechner errors and with the SM or the EM of demographics. For these estimations we use the approach of Harrison (2008) and the corresponding Stata

code.<sup>16</sup> Neither MLE model resulted in an estimated  $r = 0$  for any participant. Results can be found in the lower panel of Table 2.4.

As one can see, all estimations have a very similar mean  $r$ . Only the HL midpoint estimation technique delivers a slightly lower mean  $r$ . The estimated distributions differ mostly in their standard deviations and their extreme values. Again, the HL midpoint estimation delivers different data. In these two dimensions it is strongly different from the others. The similarity of the mean  $r$  of the five different estimation techniques combined with the support of the literature is reassuring as it suggests we analyze our data correctly. The different standard deviations and extreme values are as one would expect them.

Figure 2.2 shows scatter plots of stated probabilities vs. corrected probabilities for different correction models. The axes are organized equivalently to Figure 2.1. On the y-axis the stated probability of the participant is denoted. On the x-axis the corrected probability is denoted. We do not call it “subjective probability”, as we do not know whether this is actually the subjective probability of the respective participants. It is the subjective probability of a participant, assuming our estimation of risk aversion and our model of belief correction is correct. This figure displays all elicited beliefs of all participants of all matches and all matching groups for six different belief correction models, which gives us 11,034 points (“Market Entries”) in each scatter plot.

The graphs of Figure 2.1 serve as references in analyzing Figure 2.2. Classifications of a participant’s risk aversion can be made accordingly. A point on the identity line means that the respective belief was not corrected. A point above the identity line means that the stated probability is biased towards 1 and was corrected downwards. A point below the identity line presents a stated probability biased towards 0 which was corrected upwards. All points are displayed in the same size and at the exact location the data determines them (i.e. no “jitter” function for the graphs). This means that from the graphs we can unambiguously derive the underlying probability combination data but not the frequency of the probability combinations. 100 points on the same coordinates would look identical

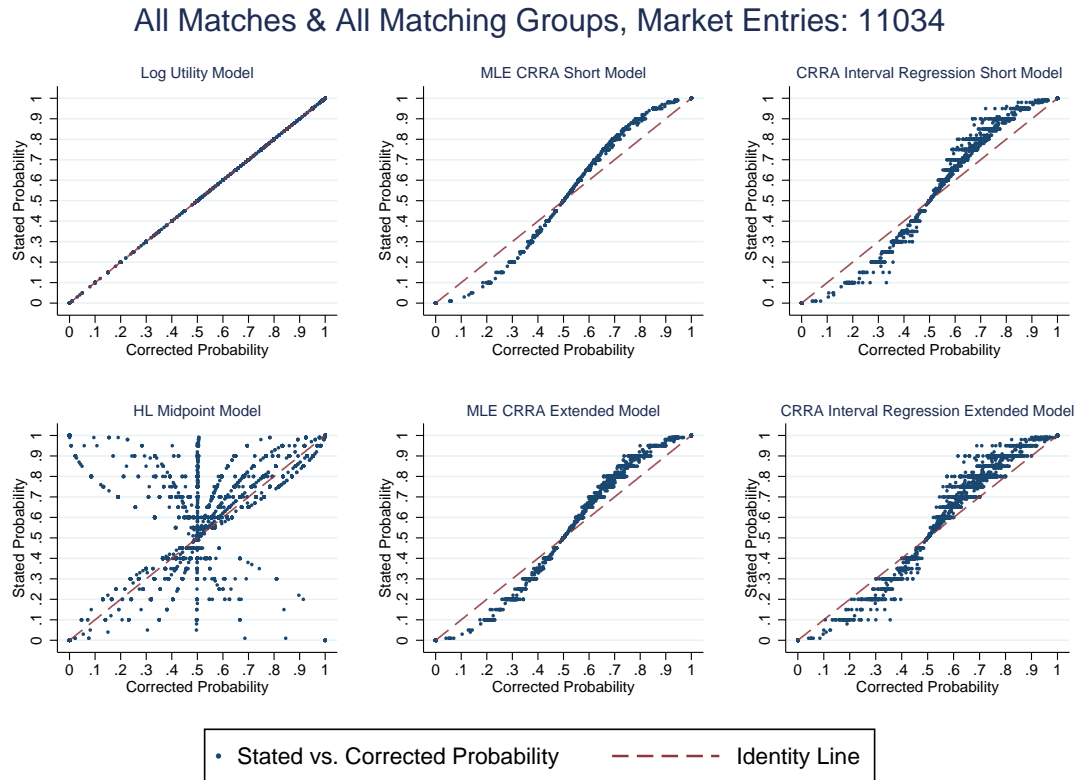
---

<sup>16</sup>We kindly thank Glenn Harrison for making the code publicly available.

## WHAT MAKES THEM TICK

to one point on the respective coordinates. This lets us easier identify the shape of the function that translates stated probabilities into corrected probabilities.

Figure 2.2: Stated Probability vs. Corrected Probability for Different Models of Risk Aversion Estimation



The upper left plot – “Log Utility Model” – plots the corrected probability against the stated probability if we assume log utility for all participants. All points are on the identity line. The lower left plot – “HL Midpoint Model” – displays the data assuming participants had levels of risk aversion determined by the HL Midpoint Model. We can identify the function shapes of very risk averse participants, slightly risk averse participants, slightly risk loving participants, and very risk loving participants. We also see many data points at 0.5 of the x-axis. These are data points derived from the risk neutral participants. Even with our assumptions made for Function 2.6 there is no clear guideline of how to correct stated beliefs  $q \notin \{0, 0.5, 1\}$ . We therefore corrected such beliefs for the 32 risk neutral participants to 0.5. Overall, there is no consistent pattern in this plot.

The two scatter plots in the middle of the upper and the lower panel denote the correction

results if we assume the risk aversion parameters estimated by the two MLE models. As expected from the results of Table 2.4 the dots together mimic the function of a slightly risk averse participant. The dots of the extended model are somewhat more dispersed than the dots of the short model but they show the same overall pattern. This pattern is also observable in the plots for both interval regression models. Compared to the extended MLE model there is more dispersion for the short model of interval regression and even more dispersion in the extended interval regression model. All these patterns are in line with the summary statistics of the  $r$  estimations.

#### 2.4.4 Comparison of Mean Beliefs, Equilibrium Prices, and Final Market Prices

The first statistic we compute is the mean of beliefs for each market and each matching group. When computing this statistic we average over the beliefs of all entered participants. We do not weight or exclude the beliefs of single participants in any way. Apart from Ottaviani and Sørensen (2012) who weight the beliefs in a way that does not apply in our context, the current theoretical literature does not weight the traders' beliefs when the mean belief is computed. There is also no distinction between active and inactive individuals. All individuals trade. In our markets, entered participants who did not trade or submit bids/asks could have done this as a response to the current prices and the bid/ask queues or because of no interest in trading at all. An analysis of participants' characteristics, beliefs and trading behavior could guide the criterion to discriminate between these participants. However, such a criterion would always have arbitrary elements. We therefore assume that beliefs are uncorrelated to the decision of active participation. This way, we use all submitted unweighted beliefs to determine the mean belief. We do this for the corrected as well as for the stated beliefs.

Our calculation of the theoretical equilibrium prices closely follows Wolfers and Zitzewitz (2007). Based on the estimation of the risk aversion parameters, we can derive the

individual demand functions. These in turn lead to the aggregated demand functions. Through the complementarity of the two securities, demand of asset 1 determines supply of asset 2. The price at which aggregated demand is equal to aggregated supply is the equilibrium price. Such a way of computing the equilibrium price demonstrates the reason of the shift of the main research interest away from the marginal trader. All traders are marginal in this setting. There is no specific marginal trader any more.

Regular prediction markets do not yield a defined final price. There are bids and asks in the respective queues. And there is a last price at which a trade was executed before the market closed. Both information sources can be used to compute the final price. Determining this price via the spread of the highest bid and lowest ask at the end of a market makes the existence of informative bids and asks necessary. Especially at the end of a thin market when (most) traders have already satisfied their trading demand, this can be difficult to fulfill. If bids and asks are not just marginally outside of the participants' trade causing interval, they are arbitrary. So as long as we do not know, whether their spread is marginally greater than the interval that leads to trade, we can not compute a meaningful price inside of the spread. Hence, we focus on the prices of executed trades.

We compute the final prices for each asset as a volume-weighted mean of the last ten percent of trades in each market. This procedure makes higher volume trades more important. We assume that these trades are less likely to be outliers or mistakes. Additionally, the final price is not tied to a certain time window before the end of the period of potential trade but focuses on the end of the active trading period. Furthermore, we assume that prices of later trades are more meaningful than prices of earlier trades, because more trades have included information in the prices. This procedure can lead to final prices of the two assets that do not sum up to 1. We correct for this by normalizing the final prices of the two assets for each market. The price of one asset is divided by the sum of both prices. Final prices of prediction markets are usually computed in such a way. Examples of not using the very last price of a market, averaging, or normalizing include Forsythe et al. (1992), Forsythe et al. (1999), and Berg et al. (2008).

This gives us four predictive statistics for each market: Mean Stated Belief (MSB), Mean Corrected Belief (MCB), Equilibrium Price (EP), and Final Price (FP).<sup>17</sup> All these statistics can be used to issue a forecast on the respective future event. Based on the theoretical literature, especially MCB and EP should be very similar. In a first step, we compare MSB, MCB, and EP. Theoretical analyses focus on the comparison of mean true beliefs versus EP. Our corrected belief in the five different models works as a proxy for the true belief. So the MCB is our proxy for the mean true belief. The MSB is included as a robustness check. We do this in order to check the results established in the literature. In the second step, we compare MCB, EP, and FP. Of main interest is the comparison to FP. The underlying question is whether the EP is actually the right statistic for the FP. If we found significant and substantial differences of EP and FP this would pose the question of using the EP in the prediction market analysis at all.

### Graphical Analysis

Figure 2.3 shows a scatter plot of the two predictive statistics MCB and EP versus the MSB. Note, that the axes of these graphs are differently structured than the ones in the figures above so we cannot compare Figure 2.3 to these figures. However, the graphs follow the presentation of the results of different models of Figure 2.2. Due to the elicitation method of the beliefs, the belief correction method, and the normalization of the final prices, the beliefs and prices of Asset 1 and Asset 2 always sum up to 1. Therefore, all plots only display the respective statistics for asset 1 (“Team of record wins”). As in most cases the participants rated the team of record as the better team, most of the probabilities are greater than 0.5.<sup>18</sup> Keep in mind that rank order findings for MSB, MCB, EP, and FP greater 0.5 for a certain asset are flipped for the complementary asset.

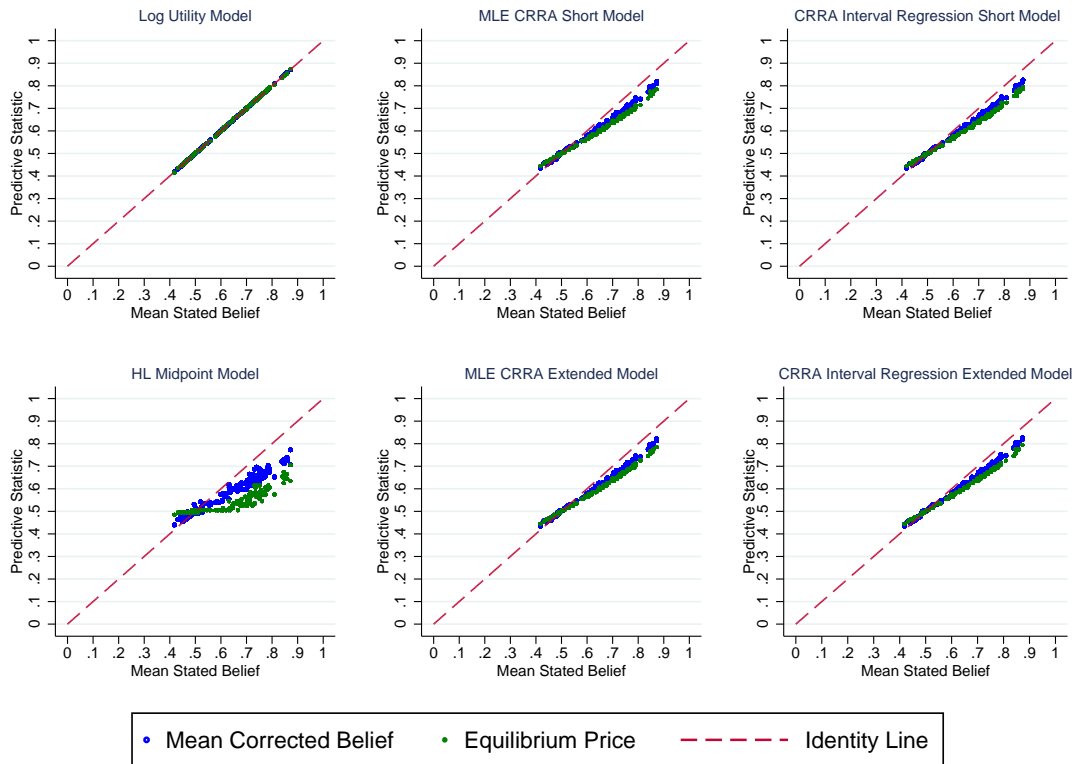
---

<sup>17</sup>Keep in mind that in our context the term “probability” of the probability correction literature has the same meaning as the term “belief” of the prediction market literature.

<sup>18</sup>This was in line with the official FIFA Ranking at <http://www.fifa.com/worldranking/rankingtable/index.html> This ranking does not get updated during a world cup, only just before and just after it and every few weeks throughout the year. The better team according to the FIFA ranking usually won the match.

## WHAT MAKES THEM TICK

Figure 2.3: Predictive Statistics vs. Mean Stated Belief for Different Models of Risk Aversion Estimation



Each data point represents one predictive statistic of each match and each matching group. The x-axis denotes the mean stated belief. The y-axis denotes the respective predictive statistic – the mean (corrected) belief (blue circles) or the equilibrium price (green dots). A dot on the identity line means that the respective statistic is equal to the mean belief. A dot above/below the identity line means that the respective statistic is greater/less than the mean stated belief.

Again, the blue dots of the upper left plot – “Log Utility Model” – show mean corrected beliefs vs. mean stated beliefs if we assume log utility for all traders. As shown in Section 2.2.4, this results in  $MSB = MCB$ . Supporting this notion, all MCB data points are on the identity line. Gjerstad (2005), Wolfers and Zitzewitz (2007), and Ottaviani and Sørensen (2012) state that with such a utility function  $MCB = EP$ . Our results are accordingly, so the green dots are displayed inside of the blue circles. The plot of the “HL Midpoint Model” shows that the data points of the MCB in most cases deviate from the



MSB. They are usually lower than the MSB for mean stated beliefs greater than 0.5 and higher for mean stated beliefs less than 0.5. The data points of the EP also show these characteristics. They additionally deviate from the MCB and are stronger pulled towards 0.5 than the MCB. This is stronger the greater the distance to an MSB of 0.5.

The plots of the four other models are strikingly similar. These plots are not four times the same plot and there actually are differences, but they are hard to find. All plots share the qualitative characteristics of the HL Midpoint Model plot. Usually, the MSB is greater or equal than the MCB which is greater or equal than the EP for an MSB greater than 0.5. This relation is flipped for an MSB less than 0.5. Usually, the MSB is less or equal than the MCB which is less or equal than the EP. Again, both differences seem to grow if we move away from an MSB of 0.5.

Summarizing the graphical analysis we can say that MSB, MCB, and EP coincide for log utility. For the other models, the graphical analysis suggests  $MSB \geq MCB \geq EP$  for  $MSB > 0.5$  and  $MSB \leq MCB \leq EP$  for  $MSB < 0.5$ . By definition of our correction method we have the following properties:  $MCB > 0.5 \iff MSB > 0.5$ ,  $MCB = 0.5 \iff MSB = 0.5$ , and  $MCB < 0.5 \iff MSB < 0.5$ . So we can state another suggestion from the graphical analysis:  $MCB \geq EP$  for  $MCB > 0.5$  and  $MCB \leq EP$  for  $MCB < 0.5$ . Which certain model the researcher employs for estimating the degree of risk aversion, seems to be of minor importance as long as the estimated  $r$  is in a realistic range. But then again, four of our models yield similar results for the estimated  $r$ . The only model in which MSB, MCB, and EP seem to be notably different is one in which we deliberately chose a very coarse measurement of the risk aversion parameter.

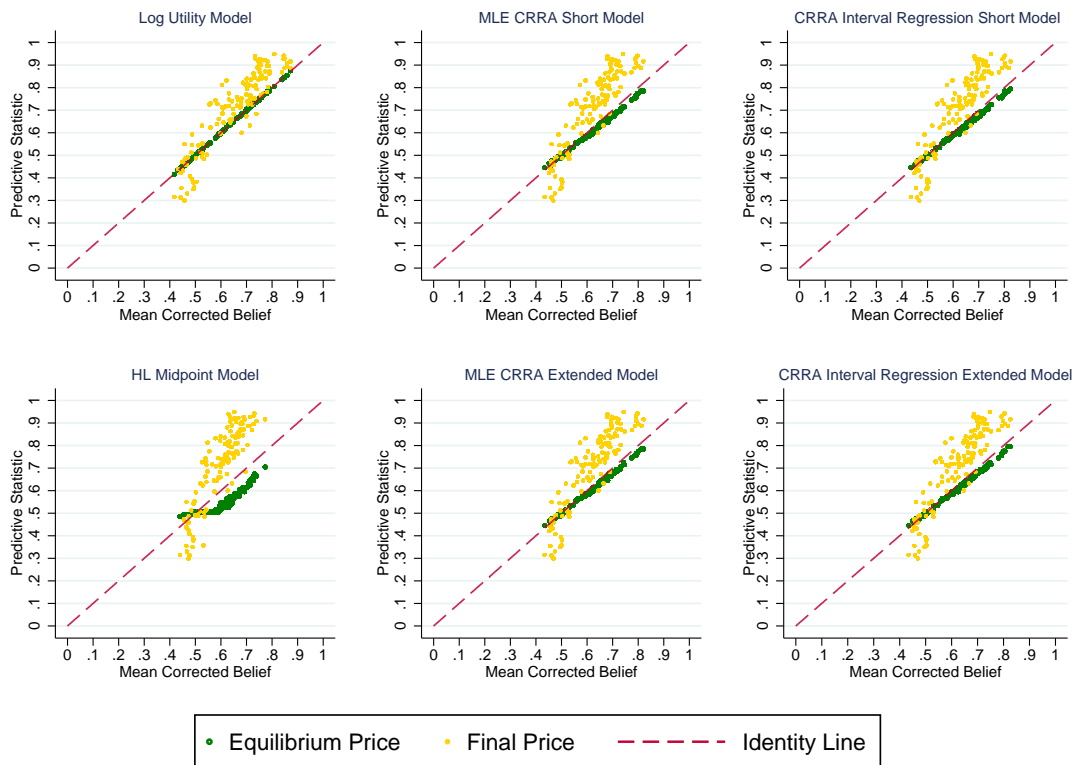
We expected the MCB to be closer to 0.5 than the MSB, because we saw that in Section 2.4.3 that most belief corrections in our dataset are towards 0.5 (because of the risk aversion of the participants). We had no hypothesis for the relation of the EP towards the MSB and the MCB apart that by definition of the correction it should be the same.

Figure 2.4 displays the predictive statistics – EP (green) and FP (yellow) – versus the

## WHAT MAKES THEM TICK

MCB for each of the six models. Note, that denoted on the x-axes is now the MCB as we do not need the MSB for discussion any more. This not only means that we cannot compare the plots to the ones in Figure 2.3, but also that we cannot compare the individual plots to each other. Each one has the MCB of the respective model denoted on its x-axis. Thanks to the information from the graphical analysis of Figure 2.3, we know that the x-values of the four MLE and interval regression models are roughly the same. A data point near the identity line denotes a similarity to the respective MCB. We therefore expect all green dots to be close to the identity line.

Figure 2.4: Predictive Statistics vs. Mean Corrected Belief for Different Models of Risk Aversion Estimation



For each plot, two properties are evident at a glance. Firstly, most of the data points for the FP are nowhere near the data points for the EP. Secondly, most of the data points for the FP are nowhere near the identity line. It seems like the EP is not a very good predictor for the FP. At further analysis for all models apart from the Log Utility Model, we can see that most of the EP data points are below the identity line for an  $MCB > 0.5$

while most of the FP data points are above the identity line for an  $MCB < 0.5$ . The relations of FP, EP, and identity line to each other are not very clear for an MCB close to 0.5. The EP dots seem to be above the identity line, but there is no clear pattern for the FP. As expected, the green dots are very close to the identity line. This again shows how similar MCB and EP are for the MLE and interval regression models.

We checked the results for several different computations of the FP. We varied the percentage of the last trades in the interval  $[0.001, 1]$  and computed the prices with and without volume weighting. Our results are robust to these checks. This is suggestive evidence for two things. Firstly, prices do not seem to be very volatile. And secondly, outlier price trades seem to have been traded at a similar volume as non-outlier price trades – if there are any outlier price trades.

### Statistical Tests

From here on, we focus on the extended interval regression model. We do not regard the Log Utility Model any more and deliver results for the four other models only if they differ qualitatively or are necessary for comprehension.

In our first tests we compare the statistics presented in Figure 2.3 – MSB, MCB, and EP – against each other. In a conservative approach, we generate the averages for each of the three statistics over all 16 matches. This gives us one independent observation per matching group, 8 overall. A Friedman Test compares the average MSB, average MCB, average EP against each other. The test rejects the null hypothesis that all the statistics are identical (Friedman test statistic: 16.0, p-value = 0.0003). The p-value is further reduced, if we do not average over the matches and use 128 observations for each statistic (Friedman test statistic: 109.23, p-value < 0.0001).

As a next step we make pairwise comparisons via the Wilcoxon-signed-rank test. The conservative approach with eight independent observations per statistic lets us reject the null hypothesis of equal average MSB and average MCB ( $z = 2.521$ , p-value = 0.0117). The comparisons of average MCB vs. average EP and average MSB vs. average EP yields

the same results with the identical test statistics. Checking the actual numbers shows that *average MSB* > *average MCB* > *average EP* in each matching group. Keep in mind that this ordering is entirely driven by the fact that most of our MSB are greater than 0.5. These relations would be flipped if we looked at the statistics for the complementary event.

We conclude that the belief correction seems to make a difference as well as the calculation of the EP. However, the quantitative differences are marginal, even if we normalize the differences by the EP. The absolute difference of EP and MCB, normalized by the EP has a mean of 0.0261, a standard deviation of 0.0138 and lies in the interval [0.0004, 0.0522]. So although this difference is statistically significant, it is indeed negligible. But the approach of the EP has two steps. So of interest are also absolute and normalized difference between the MSB and the EP. If there were no large differences either, we would not need the belief correction and the EP calculation. This difference has a mean of 0.0740, a standard deviation of 0.0363 and lies in the interval [0.0005, 0.1263].<sup>19</sup> So this difference is indeed economically significant and we should not use the MSB instead of the EP. Figure 2.3 supports this finding.

### **Result 1 [MSB vs. MCB vs. EP]**

*MSB, MCB, and EP are significantly but not substantially different to each other. For an MSB greater than 0.5, we usually have  $MSB > MCB > EP$ . This is flipped for an MSB less than 0.5.*

Our results confirm the established literature. For reasonable degrees of risk aversion and realistic distributions of beliefs<sup>20</sup>, the equilibrium price is usually very close to the mean belief (MCB in our case). But the most important question is: In what way is the equilibrium price a substitute for the final price on prediction markets? We answer

---

<sup>19</sup>The respective figures for the midpoint model are as follows. Absolute normalized difference of the MCB to EP: Mean: 0.1066, Std. Dev. : 0.0502, Range: [0.0048, 0.201]. Absolute normalized difference of the MSB to EP: Mean: 0.2137, Std. Dev. : 0.1102, Range: [0.0046, 0.407].

<sup>20</sup>By definition, our belief distributions are realistic.

this question by comparing the final prices on our prediction markets to the respective equilibrium prices. The graphical analysis suggests that there is a substantial difference between these two statistics and that the final prices are closer to the extremes than the equilibrium prices in most cases.

Again, we start conservatively and average EP and FP over all matches. For each matching group the average EP is greater than the average FP. The result of the Wilcoxon-signed-rank test ( $z = -2.521$ ,  $p\text{-value} = 0.0117$ ) lets us again reject the null hypothesis of equal values. If we do not average but use each single of the 128 observations per statistic, we again reject the null hypothesis and receive test statistics of  $z = -7.689$ ,  $p\text{-value} < 0.0001$ . We take a look at the actual numbers in order to determine the economic significance. The absolute difference of the FP and the EP, normalized by the EP has a mean of 0.1983, a standard deviation of 0.1051 and lies in the interval  $[0.0011, 0.4383]$ .<sup>21</sup> So we can conclude that the EP is substantially different to the FP. This is a crucial finding. It questions the fundament of using the EP. Additionally, we see that each step from MSB towards MCB and towards EP usually *increases* the difference to the FP.

## **Result 2 [MCB vs. EP vs. FP]**

*EP and FP are significantly and substantially different. For an MCB greater than 0.5, the EP is usually less than the MCB, while the FP is greater than the MCB. This finding is also flipped for an MCB less than 0.5.*

As a robustness check for low participation markets, we run the tests for mean stated/corrected beliefs also for median stated/corrected beliefs for all models but the Log Utility Model. We receive the same qualitative results with one exception – the equilibrium price is usually slightly greater than the median corrected belief for the MLE and interval regression models. Additionally, we run the entire set of tests without the matches of Germany. Germans were by far the greatest group of participants and therefore their

---

<sup>21</sup>The respective figures of the Midpoint Model are: Mean: 0.3465, Std. Dev.: 0.1698, Range: [0.0040, 0.7559].

support for the German team could have biased the results in either direction. Emotional hedging as well as an optimism bias was possible. Excluding the Germany matches from the tests does not alter the qualitative results in any way.

### 2.4.5 Predictive Accuracy

Measuring the predictive accuracy of a forecast for one event is usually done via the squared prediction error. It is constructed similarly to quadratic scoring rules. The forecasted (binary) event is allocated 1 if it occurred and 0 if not. The difference between the eventual outcome and the forecast is squared and the result is the squared prediction error. When we have more than one forecasted event from a forecaster, we can employ the measurement introduced by Brier (1950). It is now called the ‘‘Brier Score’’ (BS) and essentially the mean of the squared prediction error over all possible states and all events on which forecasts were issued.

Formally, the BS is calculated in the following way:  $p_{sn}$  denotes the probability the forecaster assigned to the outcome  $s \in \{1, 2, \dots, S\}$  of event (match in our case)  $n \in \{1, 2, \dots, N\}$ .  $\mathbf{1}_A(\text{condition})$  is again the indicator function.

$$BS = \frac{1}{N} \sum_{n=1}^N \sum_{s=1}^S (p_{sn} - \mathbf{1}_A(sn))^2 \in [0, 2]$$

The BS has a negative orientation, so the forecaster tries to minimize it. It is minimized if the forecaster assigns 1 to the eventual outcome of each event. It is maximized if the forecaster assigns 0 to the eventual outcome and 1 to any other outcome for each event.

For binary forecasts with  $S = 2$ , a simplified version of the BS exists:

$$BS = \frac{1}{N} \sum_{n=1}^N (p_n - \mathbf{1}_A(n))^2 \in [0, 1] \tag{2.7}$$

The respective forecast  $p_n$  must always be the forecast for the same outcome  $s = 1$  or  $s = 2$  for all  $N$  events. In a setting with  $S = 2$ , the simplified BS is a proper scoring

rule, as in such a case  $p_2$  is fully determined by  $p_1$  and vice versa. For  $S > 2$  it is not. In our setting  $S = 2$ , so we use the simplified version. The simplified BS also has a negative orientation, is minimized if the forecaster assigns 1 to the eventual outcome for all  $n$ , and maximized if the forecaster assigns 0 to the eventual outcome for all  $n$ .

In our context, one forecaster is one of the eight matching groups. Table 2.5 summarizes the BS computed for these matching groups. It lists the BS for all four predictive statistics for the set of all 16 matches.

Table 2.5: Brier Scores of the 8 Matching Groups for Different Predictive Statistics and the set of all 16 Matches

| Matching Group | BS of FP | BS of EP | BS of MCB | BS of MSB |
|----------------|----------|----------|-----------|-----------|
| 1              | 0.2030   | 0.2082   | 0.2057    | 0.2052    |
| 2              | 0.2132   | 0.2052   | 0.2029    | 0.1998    |
| 3              | 0.2056   | 0.2059   | 0.2030    | 0.2008    |
| 4              | 0.2067   | 0.2018   | 0.1986    | 0.1933    |
| 5              | 0.2046   | 0.2100   | 0.2095    | 0.2078    |
| 6              | 0.2310   | 0.2089   | 0.2084    | 0.2067    |
| 7              | 0.2193   | 0.2033   | 0.1986    | 0.1939    |
| 8              | 0.2086   | 0.2032   | 0.2017    | 0.2004    |

A Friedman test gives us the result of significant differences across all four measurements (Friedman test statistic = 13.35, p-value = 0.0039). When we compare the Brier Scores pairwise with a Wilcoxon-Signed-Rank Test, we cannot reject the null hypothesis of no differences between FP and EP ( $z = 1.4$ , p-value = 0.1614). However, we find a marginally significant difference between the FP and the MCB ( $z = 1.82$ , p-value = 0.0687), and a clearly significant difference between FP and MSB ( $z = 2.1$ , p-value = 0.0357).

From here on,  $BS(\text{statistic})$  denotes the BS of the respective statistic. Again, a Wilcoxon-Signed-Rank test detects a difference between the  $BS(EP)$  and  $BS(MCB)$  ( $z = 2.521$ , p-value = 0.0117), as well as between the  $BS(EP)$  and  $BS(MSB)$  ( $z = 2.521$ , p-value = 0.0117). Finally the exact same result is derived for a test between the  $BS(MCB)$  and  $BS(MSB)$  ( $z = 2.521$ , p-value = 0.0117). The last three results can be explained by the fact that  $BS(EP) > BS(MCB) > BS(MSB)$  for each matching group.

If we narrow our focus on the twelve matches without Germany, the picture slightly changes. Table 2.6 shows the respective BS. For each matching group, the  $BS(FP)$  is

greater than all other statistics. The picture is less clear for the other three statistics.

Table 2.6: Brier Scores of the 8 Matching Groups for Different Predictive Statistics and the subset of 12 Matches without Germany

| Matching Group | BS of FP | BS of EP | BS of MCB | BS of MSB |
|----------------|----------|----------|-----------|-----------|
| 1              | 0.2115   | 0.2103   | 0.2083    | 0.2090    |
| 2              | 0.2233   | 0.2086   | 0.2080    | 0.2052    |
| 3              | 0.2120   | 0.2066   | 0.2048    | 0.2041    |
| 4              | 0.2198   | 0.2036   | 0.2002    | 0.1959    |
| 5              | 0.2312   | 0.2145   | 0.2150    | 0.2141    |
| 6              | 0.2464   | 0.2127   | 0.2138    | 0.2120    |
| 7              | 0.2369   | 0.2031   | 0.1994    | 0.1952    |
| 8              | 0.2197   | 0.2087   | 0.2074    | 0.2080    |

The usual Friedman test is significant and shows us the differences of the four sets of BS (Friedman test statistic = 19.8, p-value = 0.0002). Pairwise comparisons with a Wilcoxon-Signed-Ranks test shows significant differences of the BS(FP) to any other BS with the identical test statistics ( $z = 2.521$ , p-value = 0.0117).

The same test rejects the null hypothesis for the comparison of BS(EP) and BS(MCB) ( $z = 1.96$ , p-value = 0.0499) as well as for the comparison of BS(EP) and BS(MSB) ( $z = 2.521$ , p-value = 0.0117). Finally, we reject the null hypothesis for the comparison of BS(MCB) and BS(MSB) ( $z = 1.96$ , p-value = 0.0499).

So the test results between BS(EP), BS(MCB), and BS(MSB) are identical for both sets of matches. However, while there are no significant differences between BS(FP) and BS(EP) in all 16 matches, the BS(FP) gets relatively worse for the subset of matches without Germany and statistically different from BS(EP). Generally, the predictive accuracy of the four predictive statistics are ordered in the following:  $BS(FP) \geq BS(EP) > BS(MCB) > BS(MSB)$ . Keep in mind that a lower value is preferable to a higher value. So to derive good predictions in events similar to ours, one does not need prediction markets. The simplest belief elicitation method and the stated beliefs deliver the best forecasts (if it attracts the same set of participants). But the differences in the predictive accuracy are only marginal. The final prices are usually more extreme than the other predictions. Hence the accuracy result is most likely to be caused by a strong impact



of high squared prediction errors when the prediction is wrong. As the squaring of a prediction error weights the prediction errors by themselves, the penalty for an extreme and wrong prediction is stronger than the one for a mediocre and wrong prediction.

**Result 3 [Predictive Accuracy]**

*The simplest predictive statistic – the mean stated belief – outperforms all other statistics in predictive accuracy. The final price has the worst predictive accuracy. The equilibrium price and the mean corrected belief rank in between the other statistics. Overall the absolute differences are only marginal.*

## 2.5 Conclusions

We analyzed prediction markets in the context of the FIFA Soccer World Cup 2006 and wanted to know in what way these markets are actually influenced by the utility functions and the beliefs of their traders. The established theoretical literature focuses on these two input factors and derives two predictive statistics, the mean of beliefs and the equilibrium price. Both are regarded as very similar and the equilibrium price serves as a substitute for the final price. If this was the case, prediction markets were nothing more but an instrument to elicit traders' beliefs. Any incentive compatible belief elicitation approach which gathered the same set of participants would be equivalent, and preferable if cheaper. However, not a single paper so far tested, whether the equilibrium price is actually a good proxy for the final price. The final price is the statistic which is used to generate the prediction market's forecast. A substantial difference between these statistics would challenge the prevalence of the equilibrium price as a substitute for the final price. The equilibrium price could still be the best proxy we have for the final price, but the difference of these statistics could be so big that it would render the equilibrium price an unreasonable instrument.

Summing up our results, it seems like the established literature was right and wrong. Although statistically significant, we only find negligible differences between the mean belief and the equilibrium price. This confirms the established results. But the literature was misguided when using the equilibrium price. In our data it is significantly and substantially different from the final price. Actually, the mean belief and even the mean uncorrected belief are usually closer to the final price than the equilibrium price. We do not find qualitative differences in these results across all realistic utility function models. This supports the general notion of only marginal impacts through utility functions as long as they are realistically estimated. Another finding of our data is that we do not see qualitative differences between the model results if we employ differently estimated degrees of risk aversion. The literature says, as long as the distribution of the risk aversion coefficient is empirically valid, we will receive similar results for different degrees of risk aversion. We do not find evidence against that.

Obviously, prediction markets do more than just eliciting beliefs and weighting them to achieve the equilibrium price. Not only risk aversion and beliefs make them tick. Dynamic factors seem to play another role. At least three explanations of our results are possible. The first one is that these markets gather information during the trading period which is by construction not reflected in our elicited beliefs. This would mean that the submitted beliefs during the runtime should change. In a quick analysis, we do not find this. Additionally, during the runtime of all markets, there was no release of information so severe that it would explain the magnitude in the difference of equilibrium price and final price. The second explanation would be that prediction markets weigh the participants' beliefs in a way that we do not fully understand yet. Tentative support for this hypothesis (and against the first one) gives the relation of the mean corrected belief and the final price towards 0.5. In only eleven out of 128 occasions, the mean corrected belief was greater than 0.5 while the final price was less, or vice versa. And in all these occasions, the mean corrected belief was very close to 0.5 – in the range  $[0.46, 0.53]$ . So there was generally a high degree of uncertainty in the respective prediction market. The

third explanation is simply that we did not elicit the true beliefs and utility functions of the participants. But this would mean that state-of-the-art methods of economic research do not result in the correct data and that our results can still be taken as the most realistic approach so far.

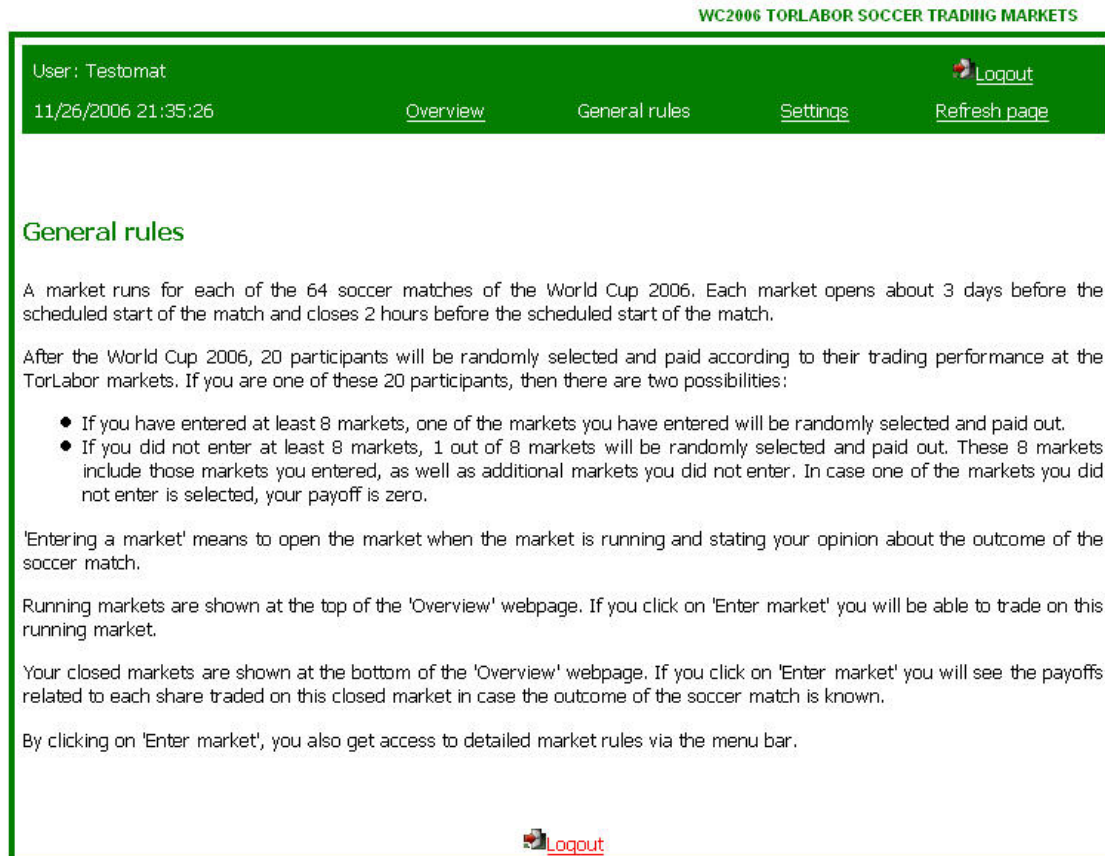
Concerning the accuracy of our predictive statistics, we find significant differences of them. The best predictor is the mean stated belief, followed by the mean corrected belief and the equilibrium price. The final price surprisingly ranks worst in our analysis of predictive accuracy. In the best setting for its performance, it is just as good as the equilibrium price. Note that the absolute differences in predictive accuracy for all four statistics are only minor. Additionally, it may still be useful to set up a prediction market to gather forecasts as this platform gathers the “right” participants.

We stress that our results have one noteworthy caveat. We elicit prior, not posterior beliefs. We therefore cannot reject any of the models presented in the literature. We do, however, offer the most extensive and realistic dataset so far. Future research should focus on dynamic aspects of final price determination and a weighted mean belief. The most prominent theoretical example of such an approach is Ottaviani and Sørensen (2012). Furthermore, the predictive accuracy should be more extensively analyzed.

## 2.6 Appendix

### 2.6.1 Screenshots



Figure 2.5: Screenshot of the General Rules Page



# WHAT MAKES THEM TICK

Figure 2.6: Screenshot of the Belief Submission Page

WC2006 TORLABOR SOCCER TRADING MARKETS

User: Testomat L16: Germany - Sweden  

11/26/2006 21:20:11

Before entering the market, please give us your opinion about the outcome of the soccer match. By doing so, you have another possibility to earn some money at TorLabor soccer trading markets.

During the registration process, you already have been offered a possibility to earn money at TorLabor soccer trading markets. Please notice that if you are rewarded for having completed this first task then you won't be rewarded for completing the present task. You will be informed only after the World Cup 2006 whether you are rewarded and if so for which task.

Later on, you will be offered a third possibility to earn money - by trading at TorLabor markets. Further details will be provided in due time. Please notice that if you are rewarded for completing the present task then you won't be rewarded for trading. You will be informed only after the World Cup 2006 whether you are rewarded and if so for which task.

**The task:**

Below this text you can see a table. The left column contains the possible outcomes of the soccer match. Please indicate in the right column, what is - according to you - the percentage chance, that each possible outcome will be the true outcome of the soccer match. These percentages that you indicate must be expressed in the form of whole numbers and the two percentages must add up to 100%. These percentages will not be revealed to any other participant.

[Examples](#) (will open in a new window)


**How is the task rewarded?**

5 randomly selected participants will be paid according to the percentages they indicated for one randomly selected soccer match. More precisely, at the end of the World Cup 2006, the three percentages that a selected participant has indicated will be rewarded in the following way:

- The selected participant will receive 100 euros multiplied by the percentage chance he/she assigned to the team that actually won the match.

[Examples](#) (will open in a new window)



| L16: Germany - Sweden       |                        |
|-----------------------------|------------------------|
| Outcome of the soccer match | Percentage chance      |
| Germany wins                | <input type="text"/> % |
| Sweden wins                 | <input type="text"/> % |



# WHAT MAKES THEM TICK

Figure 2.7: Screenshot of the Market Rules Page 1/3

WC2006 TORLABOR SOCCER TRADING MARKETS

|                     |                        |  |  |
|---------------------|------------------------|--|--|
| User: Testomat      | <b>Team A - Team B</b> |  Testomat |  Logout |
| 11/26/2006 20:42:35 | Overview               | Market rules   | Refresh page   |

OPENING TIME: 11/26/2006 21:00:00 CLOSING TIME: 11/27/2006 21:00:00

## Market rules

Here we provide the market rules for the **second round of the Soccer World Cup 2006**. Please notice that during the second round of the World Cup 2006, each soccer match between team A and team B ends up either with the victory of team A or the victory of team B.

All functions described in these market rules can be accessed via the menu at the top.

### General

The first time you enter a market, you get an endowment of 10 000 eurocents that can be used for trading in this market. Endowments as well as payoffs cannot be transferred from one market to another.

Two shares are traded on the market: the share 'Team A wins' and the share 'Team B wins'.

With your endowment of 10 000 eurocents, you can buy units of any of the two existing shares which are traded on the market. Share units which have been bought can then be sold. Buying and selling shares is not mandatory. If you do not trade at all in a market, your payoff will equal your endowment.

At the top of the overview screen, you can see your total cash as well as your available cash. Your available cash is your total cash minus the value of your current buy offers on this market. The value of a buy offer equals the quantity of the offer multiplied by the price of the offer. Details are explained in the further text.

At the bottom of the overview screen and on the page of every specific share, you can see your trades on this market so far as well as your current offers.

Your portfolio consists of the number of units held in each of the traded shares. The number of units (=quantities) are shown on the overview screen. (Below "My portfolio".)

### Market actions

Once you enter the market, you can undertake the following actions: Place a bid (an offer to buy), place an ask (an offer to sell), delete an outstanding bid or ask, buy bundles from the bank and sell bundles to the bank.

### The bank

At any point of time during the running period of a market, you can buy bundles from the bank and sell bundles to the bank. One bundle consists of one unit of each share traded in the market. The price of a bundle always equals 100 eurocents. Whenever you decide to buy bundles from the bank or sell bundles to the bank, you have to indicate the quantity.

- When buying one bundle, your owned quantity of each share increases by one.
- When selling one bundle, your owned quantity of each share decreases by one.

So, whenever you buy a bundle from the bank or sell a bundle to the bank, your portfolio changes. The quantity of bundles you can buy is restricted by your available cash. The quantity of bundles you can sell is restricted by the number of complete bundles you have in your portfolio minus your current sell offers at the market.

[Examples](#) (will open in a new window)

### The market

You can trade with other participants by submitting sell offers (asks) and buy offers (bids). Buy offers (bids) can only be submitted, if the quantity of shares you intend to buy multiplied by the intended price does not exceed your available cash. Sell offers (asks) can only be submitted if you own at least the quantity of shares you intend to trade, and the shares are not bound to other sell offers.

At the bottom of the overview-page and at the side of the page of every specific share, you can see your trades since the last login.



## WHAT MAKES THEM TICK

Figure 2.8: Screenshot of the Market Rules Page 2/3

### Placing a sell offer (ask)

A typical example for a sell offer is "I offer to sell 4 units in 'Team B wins' for 54 eurocents per unit and this offer will expire in 5 hours."

Thus, a sell offer specifies the share, a price, a number of units and optionally an expiration time. Sell offer prices must be integers and at least 1 eurocent. The number of units in a sell offer must be a positive integer.

Sell offers will only result in immediate trades if some other participant has previously submitted an offer to buy the same share at the same or a higher price. Otherwise, the sell offer is placed in a queue to await acceptance by another participant.

The 5 sell offers with the lowest prices will be posted on the market screen. If your sell offer has the lowest price of all sell offers, it will be listed directly to the left of the price the last trade occurred.

Your shares will be sold, if your sell offer is the first one in the sell offer queue and the price of your sell offer is lower or equal to the first buy offer in the buy offer queue.

- If the prices of these corresponding offers are equal, they will be executed at this price.
- If the price of your sell offer is lower than the price of the corresponding buy offer, your sell offer will be executed at the price of the buy offer. (**Example:** Your just submitted sell offer has the price 59 eurocents and the currently highest buy offer has the price of 63 eurocents. Thus, trade will take place at the price of 63 eurocents. Basically, the price of the trade is determined by the older offer.)

It may happen, that the quantity of your sell offer is higher than the quantity of the corresponding buy offer. In this case, your offer will be partially executed. This means, that as many shares as determined by the buy offer will be traded at a price determined as described above. The remaining part of your sell offer will expire.

[Examples](#) (will open in a new window)

### Placing a buy offer (bid)

A buy offer is an offer to buy in the form: "I bid 88 eurocents per unit for 8 units in 'Team A wins' and this bid will expire in 3 hours."

Thus, a buy offer specifies a share name, a price, a number of units and optionally an expiration time. Buy offer prices must be integers and at least 1 eurocent. The number of units in a buy offer must be a positive integer.

Buy offers will only result in immediate trades if some other participant has previously submitted an offer to sell the same share at the same or a lower price. Otherwise, the buy offer is placed in a queue to await acceptance by another participant.

The 5 buy offers with the highest prices will be shown on the market screen. If your buy offer has the highest price of all buy offers, it will be listed directly to the right of the last price.

The shares will be bought, if your offer is the first one in the buy queue and the price of your offer is higher or equal to the first offer in the queue for the sell offers.

- If the prices of these corresponding offers are equal, they will be executed at this price.
- If the price of your buy offer is higher than the price of the corresponding sell offer, your offer will be executed at the price of the sell offer. (**Example:** Your just submitted buy offer has the price 26 eurocents and the currently lowest sell offer has the price of 21 eurocents. Thus, trade will take place at the price of 21 eurocents. Basically, the price of the trade is determined by the older offer.)

It may happen, that the quantity of your buy offer is higher than the quantity of the corresponding sell offer. In this case, your offer will be partially executed. This means, that as many shares as determined by the sell offer will be traded at a price determined as described above. The remaining part of your buy offer will expire.

[Examples](#) (will open in a new window)

Figure 2.9: Screenshot of the Market Rules Page 3/3

### Delete a buy offer or a sell offer

On the market overview page and also on the pages of each share you see a list of your outstanding buy offers and sell offers. At any time any of these offers may be deleted.

Note that outstanding offers cannot be revised, but outstanding offers can be deleted and new offers submitted. Offers cannot be deleted after they have been accepted by another participant.

### Buy offer and sell offer queues

When buy offers and sell offers are submitted by participants, they are placed in the "buy offer queue" and the "sell offer queue", respectively. Each queue is ordered according to price and time of issuance of the offers; if two or more offers at the same price appear in a queue, they are entered by time with older offers appearing ahead of newer offers.

The prices and quantities displayed you when you log into the market are the 5 highest buy offer prices in the buy offer queue and the 5 lowest sell offer prices in the sell offer queue. In the center you see the price at which the last trade occurred. Has there been no trade yet, you will see "-".

Offers remain in the queues until they are deleted by the participant who issued them, they expire or they are accepted by another participant and result in a trade.

### Trade with yourself

Trading own offers is allowed. However, it will not affect the published last price and will not be listed as a trade. It can be viewed as taking back part of your own offer.

**How are your payoffs determined?** At the end of the match, the share that corresponds to the outcome of the match will pay 100 eurocents, all other shares will expire worthless. All non-executed offers will also expire without trade.

Therefore, your payoffs correspond to the sum of

1. 100 eurocents multiplied by the number of your owned share that refers to the outcome of the match, and
2. your total cash.

**Example:** Consider the market to the match Team A - Team B. Imagine the result is 3:1 for Team A.

You own 47 shares of "Team A wins" and 399 shares of "Team B wins". Your total cash is 7 292 eurocents.

Therefore your payoff is computed by:

$47 * 100 = 4\ 700$  eurocents due to your "Team A wins"-shares

additionally

$7\ 292$  eurocents due to your total cash.

Yielding to a payoff of **11 992** eurocents.

**Please remember that after the World Cup 2006, 20 participants will be randomly selected and paid their payoffs in one randomly selected market. Thus, this market is only paid if you are among the 20 randomly selected participants and this market is the randomly selected market. For details see the general rules.**





# WHAT MAKES THEM TICK

Figure 2.10: Screenshot of a Market

**WC2006 TORLABOR SOCCER TRADING MARKETS**

User: Testomat [Exit market](#) [Logout](#)

**L16: Germany - Sweden**

11/27/2006 01:12:21 [Overview](#) [Buy/sell bundles](#) [Market rules](#) [Refresh page](#)

OPENING TIME: 11/26/2006 21:15:00 CLOSING TIME: 11/27/2006 21:15:00 **To update the information, refresh**

Your total cash: 8400 eurocents \*Your "available cash" is your total cash minus the value of your current buy offers on this market.  
 Your available cash\*: 7194 eurocents

| Share        | My portfolio | Sell offers (asks) |  |  |  | Last price | Buy offers (bids) |  |  |  |       |                      |
|--------------|--------------|--------------------|--|--|--|------------|-------------------|--|--|--|-------|----------------------|
| Germany wins | 16           | Price              |  |  |  | -          | 67                |  |  |  | Price | <a href="#">Buy/</a> |
|              |              | Qnt.               |  |  |  |            | 18                |  |  |  | Qnt.  | <a href="#">Sell</a> |
| Sweden wins  | 16           | Price              |  |  |  | -          |                   |  |  |  | Price | <a href="#">Buy/</a> |
|              |              | Qnt.               |  |  |  |            |                   |  |  |  | Qnt.  | <a href="#">Sell</a> |

My current offers

| Share        | Bid type | Price | Quantity | Expiration date     |                        |
|--------------|----------|-------|----------|---------------------|------------------------|
| Germany wins | Buy      | 67    | 18       | 11/27/2006 03:41:39 | <a href="#">Delete</a> |

My trades in this market

| Date and time | Share | Bid type | Price | Quantity |
|---------------|-------|----------|-------|----------|
|               |       |          |       |          |

[Logout](#)

Notes: This screenshot shows an actual market. As an example we picked the match Germany vs. Sweden in the round of the last 16 teams. On the very top, trader participant found browsing links. Just below this area, general information on the respective market and the participant's financial situation are given. In the center of the page, the participant found rows of information on each of the "shares". The participant's holding of the respective share is shown, queued bids and asks as well as the last price to which a share was traded. In the lower part of the screen, the participant's current offers as well as his trading history were displayed.

# WHAT MAKES THEM TICK

Figure 2.11: Screenshot of the Offer Submission Page

**WC2006 TORLABOR SOCCER TRADING MARKETS**

User: Testomat [Exit market](#) [Logout](#)

11/27/2006 01:18:12 [Overview](#) [Buy/sell bundles](#) [Market rules](#) [Refresh page](#)

OPENING TIME: 11/26/2006 21:15:00 CLOSING TIME: 11/27/2006 21:15:00

Your total cash: 8400 eurocents \*Your "available cash" is your total cash minus the value of your current buy offers on this market.  
 Your available cash\*: 7194 eurocents

| Share        | My holdings | Sell offers (asks) |  |  |  | Last price | Buy offers (bids) |  |  |  |       |
|--------------|-------------|--------------------|--|--|--|------------|-------------------|--|--|--|-------|
| Germany wins | 16          | Price              |  |  |  | -          | 67                |  |  |  | Price |
|              |             | Qnt.               |  |  |  |            | 18                |  |  |  | Qnt.  |

Submit offer for: "Germany wins"

Buy/ Sell:

Price (per share):

Quantity:

expires in hours:

My current offers

| Bid type | Price | Quantity | Expiration date     |                        |
|----------|-------|----------|---------------------|------------------------|
| Buy      | 67    | 18       | 11/27/2006 03:41:39 | <a href="#">Delete</a> |

My trades in this market

| Date and time | Bid type | Price | Quantity |
|---------------|----------|-------|----------|
|               |          |       |          |
|               |          |       |          |

[Logout](#)

Notes: The participant was guided to this screen if he clicked on "Buy/Sell" in the market overview page. He could specify the relevant data for the offer. All information on the respective share was still displayed.

# WHAT MAKES THEM TICK

Figure 2.12: Screenshot of the Bundle Trade Page

WC2006 TORLABOR SOCCER TRADING MARKETS

User: Testomat      **L16: Germany - Sweden**      [Exit market](#)      [Logout](#)  
11/26/2006 21:42:39      [Overview](#)      [Buy/sell bundles](#)      [Market rules](#)      [Refresh page](#)

OPENING TIME: 11/26/2006 21:15:00    CLOSING TIME: 11/27/2006 21:15:00

**Your total cash: 10000 eurocents**  
**Your available cash: 8794 eurocents**

**Buy/sell bundles:**

By getting a bundle at the bank you will buy **one** share of **each kind** of shares in the market. A **complete** bundle can be resold to the bank at any time.  
One bundle costs/yields **100 eurocents**.

Buy/Sell:

Quantity:

[Logout](#)

Notes: The participant was guided to this screen if he clicked on “Buy/Sell Bundles” in the market overview page.

# Chapter 3

## Social Interaction and Information in a Relative Payment Scheme<sup>0</sup>

### 3.1 Introduction

Relative payment schemes are a useful tool to reward workers. The company knows its expenditures before the actual payment and is unaffected by positive production shocks. The worker in turn is unaffected by negative production shocks. But he is paid through a payment scheme, which is usually met with discomfort. A relative payment scheme is also susceptible to collusion. If workers find a coordination device, they could collectively lower their effort without changing their payment level. Although relative payments like rank-order tournaments have beneficial characteristics in theory (Lazear and Rosen, 1981), they are hardly used in compensating workers apart from promotion (Lazear, 1989). The potential for collusion might be a reason for this.<sup>1</sup>

An exception from the usual abstention from relative wage payments is described in Bandiera et al. (2005) for fruit pickers in England.<sup>2</sup> Our research closely focuses on

---

<sup>0</sup>This chapter is based on joint work with Alain Cohn, Ernst Fehr, and Michel Maréchal.

<sup>1</sup>An extensive overview of compensation schemes is found in Prendergast (1999). Lazear and Shaw (2007) give a broader discussion on the research agenda in personnel economics. An overview on field experiments in labor economics is given in List and Rasul (2011).

<sup>2</sup>The only field study which researches a similar payment scheme and we are aware of is Knoeber and

this paper. The authors analyze fruit pickers in England and find that a relative payment scheme in which the piece rate is determined by the performance of the individual relative to the average performance of the reference group gives lower incentives than an exogenously determined piece rate. They conclude that this is due to the workers partially internalizing the negative externality they impose on coworkers. Such internalization only happens if workers pick certain fruits which allows them to interact socially and to monitor each other. It is unclear whether this result is driven by the information exchange or the social interaction. Our research tries to shed light on this issue.

We design a natural field experiment (Harrison and List, 2004) in order to find out whether social interaction or information drove the stated results of Bandiera et al.'s (2005) study. Our participants worked in dyads and were paid through a relative payment scheme. Their task was to register library books in a database. We binary vary of social interaction and the information exchange. Our results support the notion that social interaction reduces both output level and output difference of the workers. Performance information exchange in a dyad does not lead to a significant output reduction but to a reduction of the output difference. This leads to the conclusion, that the effect in Bandiera et al. (2005) was mainly driven by the social interaction channel.

Experimental examples of research on communication as a coordination device in the context of incentive schemes include Sutter and Strassmair (2009) who have a tournament setup in which teams compete against each other. They find that communication within teams increases team efforts and communication between teams decreases team efforts. Harbring (2006) finds evidence for cooperation through communication in two different incentive schemes. Communication increases effort in a team incentives setup, whereas communication decreases effort in a tournament incentives setup. Our setting differs from these studies in the way that we want to stress the social interaction channel and therefore

---

Thurman (1994) for broiler production in the USA. In this paper, a producer's payment for each pound of broiler is inversely related to the respective producer's production costs relative to the average production costs. They find that the payment difference, not the payment level, drive performance. Furthermore, they find that handicapped competitors implement riskier strategies than others and that tournament organizers want to minimize ability differences of the players.

tie communication to social interaction.

Information as a coordination device can mainly be found in the domain of industrial organization, namely as a fostering element of tacit collusion. In field data of the credit card market, Knittel and Stango (2003) find that price ceilings can lead to tacit collusion. Such price ceilings provide a focal point on which competing firms can coordinate. An empirical study which shows tacit collusion without a clear focal point is Ellison and Wolfram (2006). In a pharmaceutical pricing setting, the companies coordinate on a certain price increase. Laboratory studies include Cason (1994), who finds that under certain conditions in a market experiment, players reveal information in order to achieve a higher joint profit. Cason and Mason (1999) stress the importance of information revelation in order to achieve tacit collusion in duopoly markets. Bajari and Hortaçsu (2004) summarize the results from research on Internet auctions and also discuss tacit collusion in this context. To the best of our knowledge, tacit collusion through the information channel was not yet explicitly analyzed in the setting of a relative payment scheme. We do not find support for information as a trigger for tacit collusion in such a context.

We also contribute to the literature on office arrangements. This strand of research mostly deals with the employees' well-being and job-satisfaction. Marans and Spreckelmeyer (1982) provide a summary of comparisons of conventional and open office designs with respect to work station satisfaction. The allotted space is the driving force in this context. Zahn (1991) analyze the impact of spatial distance of employees on their face-to-face communication. Studies on office arrangements and employees' output are rare, although these arrangements gain importance with the prevailing reduction of office space for each employee.<sup>3</sup> Our research sheds light on the incentive effects of office arrangements and information channels, when companies implement a relative payment scheme. Our supplementary data from a questionnaire on work environments, information exchange, and payment schemes at the workplace of more than 100 German establishments suggests

---

<sup>3</sup>“The average amount of space allotted to each employee shrank from 500 square feet in the 1970s to 200 square feet in 2010.” (Peter Miscovich, managing director of corporate solutions for Jones Lang LaSalle in New York.)

that companies take into account the interaction of these factors.

## 3.2 Research Design

German libraries usually organized their inventories in a classification type they had freely chosen. In order to align the organizing system for books nationwide, an increasing number of libraries follow the “Regensburger Verbundklassifikation” (Regensburg classification), introduced in 1964. In our experimental setting, participants had to enter the information for these catalogues for books of a German library. We were provided with a couple of rooms in which we set up work environments with notebooks. The experiment was conducted from May 2010 until May 2011.

### Recruitment

A few weeks before the experiment, workers were recruited to catalogue the books of the library. An advertisement was placed on a job search website and posters were hung up in student dorms. The job advertisement stated that the library was looking for reliable and independent manpower with a quick grasp of the electronic entry of books. The job advertisement described a one-time job limited to exactly four hours. No information on the wage was given. Applicants could show interest in the job via an online application form. The application form requested contact information and disposability. In addition, applicants were asked to complete an online typing speed test, which lasted five minutes. The online test was an exact reproduction of the task on the job and serves as a measure for workers’ ability. However, applicants were hired on the basis of disposability and not ability. Hired applicants received a confirmation email reminding them of the terms of work and indicating the date of their shift. A reminder email was sent out two days ahead of their shift and a reminder call was made on the evening before their shift. Applicants with insufficient local language skills were excluded from the applicant pool. All workers were unaware of participating in an experiment.<sup>4</sup>

---

<sup>4</sup>After the experiment, we asked the participants to sign a permission to use the generated data for

## Work

Workers were welcomed by one or two research assistants who explained the task and accompanied them to an office.<sup>5</sup> The workers tested the input mask under the supervision of the research assistants. Their task was to record the library books into an electronic database through an input mask. Seven packing cases were assembled in a line beside the table the participants worked at. Overall about 500 books were stored in these cases. There was no clear assignment of participants to certain boxes. For each book they had to enter its title, author(s), publisher, ISBN number and year of publication. Each participant received instant feedback about his own performance. A screenshot of the input mask can be found in the appendix.

## Treatments

We invited the workers in groups of four. If workers showed up in an odd number, they were assigned to a treatment which is left for future research. If only two workers showed up, they were assigned to one of the “Sameroom” treatments explained below. Groups of four were randomly allocated in teams of two and assigned to one of four treatments.<sup>6</sup> These treatments differed in two dimensions. The first one was information. We either provide both participants of one team with information about the other participant’s performance in real time, or not (Info/NoInfo). The other dimension was the room setting. Participants of one team either sat in the same room or in two different rooms (Sameroom/Mixedroom). If one group was split across two rooms, another group was split up as well and its members were seated in the same two rooms. This way, we kept the number of participants in one room constant. This provided us with four treatments: Sameroom/NoInfo (SN), Sameroom/Info (SI), Mixedroom/NoInfo (MN), Mixedroom/Info (MI). Please note, that workers in the Sameroom/NoInfo treat-

---

research. Participants who did not sign this permission were paid like all other participants but we do not include their data in the analysis.

<sup>5</sup>The research assistants were instructed to strictly follow a plan of procedures which included a detailed description of the communication with the workers. One research assistant cared for two workers at most.

<sup>6</sup>The research assistants made sure that the four workers were of same sex and that siblings and friends were assigned to the same treatment.



ment still knew roughly about their teammate’s performance as they sat face to face at one table. Additionally, all participants were told their performance in the online test in front of everyone. So the participants had a fair estimation about their ability relative to their team members.

The payment scheme was identical for all these treatments and was exactly described at the beginning of the session. We implemented a relative payment scheme. Workers earned 44.00 € plus a performance-based element. This element depended on the difference in logged books among team members: workers received (lost) 0.20 € for each book they logged more (less) than their team member. Input quality was not part of the payment scheme.<sup>7</sup> However, this was never mentioned explicitly in order to keep the workers from logging in mere nonsense. Workers were provided a handout with a detailed illustration of this incentive scheme. A translation of this handout can be found in the appendix. A payment scheme of this kind is usually not implemented in companies, but it covers the essential features of typical relative payment schemes. It is a zero-sum game and effort is - *ceteris paribus* - translated into (expected) higher payment.

Workers were left alone on the job. If they needed help, they could call a research assistant.<sup>8</sup> After four hours of work, the research assistants went back to the offices to calculate the workers’ pay by counting the entries of the two group members. Thereafter, workers filled in a short questionnaire, signed a form and were cashed out.

### 3.3 Theoretical Model

For clear-cut hypotheses, we have to derive a model and make assumptions on the utility function as well as the cost function of the agents.<sup>9</sup> In our model we have 2 agents:  $i = 1, 2$ . They receive relative performance incentives  $w_i = \gamma + \delta(e_i - e_j)$ , with  $\delta > 0$ . In

---

<sup>7</sup>It would have been prohibitively costly to program a routine that recognized the entered book to regard quality in a correct way.

<sup>8</sup>In the meantime, the research assistants recorded whether workers understood the task and incentive scheme instructions.

<sup>9</sup>The basic model was developed by Florian Ederer and we kindly thank him for his support.

our notation,  $j$  means “not”  $i$ . Their effort has convex costs  $c_i(e_i) = \frac{a_i}{2}e_i^2$ , where  $a_i > 0$  captures ability. So a lower  $a_i$  means a higher ability and therefore lower effort costs for the same effort. In our context, agent  $i$  is the one with the higher ability.

It is possible to implement punishment  $P$  for the agent deviating from an informal contract:  $P = 0$  without social interaction and without information,  $P > 0$  with social interaction or with information. This punishment mechanism works through social punishment. The punishing agent verbally criticizes the deviating agent as soon as the deviation from the informal contract is detected. We assume that punishment is equally high for both team members. Furthermore, we assume costless punishment like reprehension. (Cooper and Kühn (2011) show that verbal punishment can have an effect in a laboratory experiment on collusion.) So the punishment can be implemented, as soon as the workers interact face-to-face. This can be during the task in the Same Room treatments or after the task in the Mixed Room treatments. No worker knows about the performance of the other worker during the task in the Mixed Room, No Information treatment. Hence, we assume that punishment is lowest in this treatment and set it to 0. We assume risk-neutrality for both agents:  $U_i = w_i - c_i(e_i) - P$ . We also assume the absence of side payments.

### 3.3.1 Mixed Room, No Information

The agent solves the following maximization problem:

$$\max_{e_i} \left[ \gamma + \delta(e_i - e_j) - \frac{a_i}{2}e_i^2 - P \right]$$

The first-order condition yields

$$e_i^{*,MN} = \frac{\delta}{a_i}$$

and with  $P = 0$  the agent's surplus is

$$U_i = \gamma + \frac{\delta^2}{2a_i} - \frac{\delta^2}{a_j}$$

### 3.3.2 Same Room, (No) Information

With social interaction in one room the agents establish an informal contract which amounts to

$$\max_{e_i, e_j} [U_i + U_j] = \max_{e_i, e_j} [w_i + w_j - c_i(e_i) - c_j(e_j)]$$

subject to

$$w_i^* - c_i(e_i^*) \geq \max_{e_i} [w_i - c(e_i) - P] \text{ for } i = 1, 2$$

Consider the first unconstrained maximization problem which is

$$\max_{e_i, e_j} \left[ 2\gamma + \delta(e_i - e_j) + \delta(e_j - e_i) - \frac{a_i}{2}e_i^2 - \frac{a_j}{2}e_j^2 \right]$$

which has a corner solution given by

$$e_i = 0 \text{ for } i = 1, 2$$

and each agent earns a surplus given by

$$U_i = \gamma$$

The reneging constraint has to hold which yields

$$P \geq \frac{\delta^2}{2a_i} \text{ for } i = 1, 2$$

Here we can introduce differences in the Information and the No Information version of the treatments. In the Same Room, Information treatment, both workers are perfectly

## SOCIAL INTERACTION

informed about the team member's performance at any point in time. One member can implement the punishment  $P$  immediately after he realizes the other member's deviation from shirking. In the Same Room, No Information treatment he cannot identify such a deviation as clearly. So it can take time until the social punishment is implemented. To catch this, we include the discount factor  $d_h \in (0, 1)$  for  $P$  in the Same Room, No Information treatment. The subscript  $h$  is for "high". A following section introduces another discount factor so that we have to distinguish these factors.

The renegeing constraint in the Same Room, No Information Treatment is

$$d_h P \geq \frac{\delta^2}{2a_i} \text{ for } i = 1, 2$$

For the Same Room, Information treatment: If  $P < \frac{\delta^2}{2a_i}$  full collusion is no longer possible, but some degree of collusion is still feasible and the renegeing constraint will be binding.

Thus, we have

$$\gamma + \delta (e_i - e_j) - \frac{a_i}{2} e_i^2 \geq \gamma + \frac{\delta^2}{2a_i} - \delta e_j - P$$

or

$$\delta e_i - \frac{a_i}{2} e_i^2 \geq \frac{\delta^2}{2a_i} - P$$

Solving the quadratic equation yields

$$e_i = \frac{\delta}{a_i} - \sqrt{\frac{2P}{a_i}}$$

Thus, equilibrium effort is given by

$$e_i^{*,SI} = \max \left\{ 0, \frac{\delta}{a_i} - \sqrt{\frac{2P}{a_i}} \right\}$$

Equilibrium effort for the Same Room, No Information treatment is given by

$$e_i^{*,SN} = \max \left\{ 0, \frac{\delta}{a_i} - \sqrt{\frac{2d_h P}{a_i}} \right\}$$

### 3.3.3 Mixed Room, Information

Here, we introduce another discount factor  $d_l \in (0, 1)$ .  $d_l < d_h$  because the punishment can only be implemented after the task. (“l” is for “low”.) Agents establish the same informal contract like in both Same Room Treatments. The entire analysis is identical to these treatments, but instead of  $d_h$  of the Same Room, No Information treatment, we employ  $d_l$ .

This yields an equilibrium effort of

$$e_i = \max \left\{ 0, \frac{\delta}{a_i} - \sqrt{\frac{2d_l P}{a_i}} \right\}$$

### 3.3.4 Discussion

When comparing effort in the treatments of participants in the four treatments, we have the following equilibrium efforts:<sup>10</sup>

$$\begin{aligned} e_i^{MN} &= \frac{\delta}{a_i} \\ e_i^{SI} &= \max \left\{ 0, \frac{\delta}{a_i} - \sqrt{\frac{2P}{a_i}} \right\} \\ e_i^{SN} &= \max \left\{ 0, \frac{\delta}{a_i} - \sqrt{\frac{2d_h P}{a_i}} \right\} \\ e_i^{MI} &= \max \left\{ 0, \frac{\delta}{a_i} - \sqrt{\frac{2d_l P}{a_i}} \right\} \end{aligned}$$

---

<sup>10</sup>We drop the superscript \* from now on

## SOCIAL INTERACTION

Given our assumptions, the efforts are ranked as follows:  $e_i^{MN} > e_i^{MI} \geq e_i^{SN} \geq e_i^{SI}$ . We also note that the equilibrium effort (agent surplus) is decreasing (increasing) in the size of  $P$ .

When agents differ in their ability (i.e.  $a_i \neq a_j$ ), the difference in equilibrium effort decreases with the size of the punishment. For a given punishment  $P$  that is lower than the thresholds for full collusion, consider the difference in effort choices when  $a_i \neq a_j$ . We have:

$$e_i^{MN} - e_j^{MN} = \frac{\delta}{a_i} - \frac{\delta}{a_j}$$

$$\begin{aligned} e_i^{SI} - e_j^{SI} &= \left( \frac{\delta}{a_i} - \sqrt{\frac{2P}{a_j}} \right) - \left( \frac{\delta}{a_j} - \sqrt{\frac{2P}{a_j}} \right) \\ &= \left( \frac{\delta}{a_i} - \frac{\delta}{a_j} \right) - \left( \sqrt{\frac{2P}{a_i}} - \sqrt{\frac{2P}{a_j}} \right) \end{aligned}$$

We derive the differences in the SN and MI treatments accordingly:

$$e_i^{SN} - e_j^{SN} = \left( \frac{\delta}{a_i} - \frac{\delta}{a_j} \right) - \left( \sqrt{\frac{2d_h P}{a_i}} - \sqrt{\frac{2d_h P}{a_j}} \right)$$

$$e_i^{MI} - e_j^{MI} = \left( \frac{\delta}{a_i} - \frac{\delta}{a_j} \right) - \left( \sqrt{\frac{2d_l P}{a_i}} - \sqrt{\frac{2d_l P}{a_j}} \right)$$

So the effort differences in the four treatments are ordered in the same way the effort levels are ordered. This ordering is not as clear any more, if  $P$  is just as high as necessary to implement zero effort of  $j$  only. We assume that this is never the case. This is also shown in our data. No participant exerted an effort of 0.

We can also analyze which agent has more to gain from cooperation and collusion when agents are different. This can be done by looking at the full coordination/collusion reneging constraint which is given by

$$P \geq \frac{\delta^2}{2a_i}$$

Thus with  $a_i < a_j$  the more productive agent  $i$  has a higher temptation to renege on the agreement, or equivalently he has less to gain from cooperation/collusion than the less productive agent. When an informal agreement is made, the more productive agent will have to reduce his effort relative to the first best, for example when

$$\frac{\delta^2}{2a_i} \geq P \geq \frac{\delta^2}{2a_j}$$

while the other agent  $j$  will exert the effort of a setting without informal agreements.

### 3.4 Hypotheses

Our two treatment dimensions are information and social interaction. We expect that social interaction reduces the workers' output and as well as the output difference through informal contracts. We expect that information yields similar results. Workers set up similar informal contracts in all treatments apart from MN.

If the two workers of one team sit together in an office, they can interact socially. In the SI and SN treatments, a team dyad maximizes its overall payoff. This means that its members lower their output compared to a situation without social interaction and information as they will get punished, if they do not. They lower their output nonproportionally. Hence the output difference will be reduced as well. As long as the renegeing constraint holds, the informal contract stays in effect.

Spatially separated team members cannot interact socially during the task. Without information on the team member's output, this implies that selfish workers maximize their private surplus. They do not take a social punishment into account. Therefore, workers enter books until their marginal cost equals their marginal benefit of 0.20 €. However, information on the team member's performance can suffice to set up an informal contract. In this case, we expect a reduced output and a reduced output difference in treatment MI like in treatments SI and SN.

Summing up and following out theoretical analysis, we expect social interaction and information to lead to a comparatively lower output as well as a lower output difference. Differences in effort costs should prevent the output level and the output difference to achieve zero. Differences in the discount factor of the punishment lead to different magnitudes of the effect in the three treatments with social punishment.

We derive two Hypotheses:

**Hypothesis 1 [Output]**

*We expect the highest effort level in the MN treatment. Participants in the MI treatment exhibit a lower effort. The effort in the SN treatment should be lower or equal than the effort in the MI treatment. Finally, the effort in the SI treatment should be lower or equal than in the SN treatment.*

**Hypothesis 2 [Output Difference]**

*The ranking of the effort level differences is completely analogue to the ranking in Hypothesis 1. We expect the highest effort level difference in the MN treatment. Participants in the MI treatment exhibit a lower effort level difference. The effort level difference in the SN treatment should be lower or equal than the effort in the MI treatment. Finally, the effort level difference in the SI treatment should be lower or equal than in the SN treatment.*

## 3.5 Results

### 3.5.1 Descriptives

An overview on the descriptives is given in Table 3.1. We had 84 participants, whose age mostly lay in the twenties but went up to 53 years. Most of our participants were female. The number of characters in the online test - which we use as a proxy for ability - has an



## SOCIAL INTERACTION

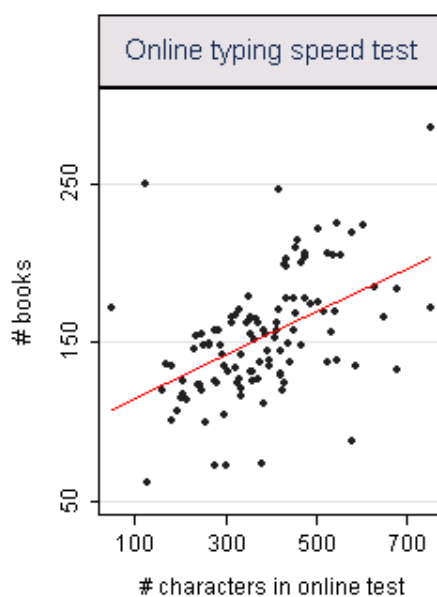
Table 3.1: Descriptives

|  |              |
|--|--------------|
| Number of Participants                     | 84           |
| Average Age (Std. Dev.)                    | 25 (6.5)     |
| Sex Female                                 | 62 %         |
| Avg. Characters in Online Test (Std. Dev.) | 388 (145.38) |

average of just below 400 and has a standard deviation of 145.

Figure 3.1 shows the correlation of the ability elicited in the online test (Number of characters) to the output in the 4 hours work time (number of books). We can see that the performance in the online test is a fair predictor of output in the work time. We will therefore use this performance as a proxy for ability of the workers. In this graph we can also see that the ability of the workers is broadly distributed.

Figure 3.1: Ability Check



Notes: The graph plots the number of books versus the number of characters in the online test for each individual.

### 3.5.2 Regressions

We pool the data and employ dummies for the treatments as well as a variable for the workers' ability and the ability difference in the output and output difference regression,

respectively. The omitted category is MN (Mixedroom/NoInfo).

### Output

Table 3.2 shows the results of the regression of output. It lists the coefficients for a linear regression of output (books) on the four treatment dummies with robust standard errors and clusters for participant groups that worked at the same time. In the second model, the ability of participants is included. Following our first hypothesis, we expect the coefficient of the “Info”-dummy (for the information dimension) to be less than 0, the coefficient of the “Same”-dummy (for the room dimension) to be less or equal than the coefficient of the “Info”-dummy, and the sum of the coefficients of both dummies and the interaction effect dummy to be at most as big as the other dummies.

First of all, we do not find a significant effect of the Info-dummy. The Same-dummy is three times the Info-dummy and significant. The interaction term is not significant. And neither is the entire model, as its entire p-value is 0.116. When we focus on the model with ability, the results are sharpened and the entire model is now significant. Ability – the number of characters in the online test – is positive and highly significant. The coefficient of the Info-dummy only changed marginally and is still insignificant. The effect of the same room is increased and is highly significant like in the other model. The interaction effect of the same room and information is increased, but is still not significant.

When we compare the coefficients of the model with ability, we can partly rank them according to our hypothesis. The omitted category (MN) indeed is the one which generates the highest output. However, the coefficient for the “Info”-dummy is not significantly different from 0, although it is negative. The coefficient for the “Same”-dummy is negative, significantly different from 0, and a post-estimation test reveals that “Same” is just significantly different from the “Info”-dummy (F-statistic = 2.86, p-value = 0.0995). Summing up all dummies yields a mediocre effect, although the “Info”-dummy is not significant. All dummies together are significantly different from 0 (F-statistic = 3.44, p-value = 0.0719). However, the information dummy jointly with the interaction dummy are not significantly

## SOCIAL INTERACTION

Table 3.2: Regression of Output on Treatment Dummies and Ability

|                | Coeff.<br>(Rob. SE)    | Coeff.<br>(Rob. SE)    |
|----------------|------------------------|------------------------|
| Info           | -6.283<br>(17.921)     | -9.472<br>(15.303)     |
| Same           | -28.100**<br>(12.543)  | -32.836***<br>(8.585)  |
| Same*Info      | 5.383<br>(23.611)      | 13.486<br>(21.002)     |
| Ability        |                        | 0.137***<br>(0.038)    |
| Constant       | 169.950***<br>(10.670) | 118.957***<br>(16.490) |
| Obs.           | 84                     | 84                     |
| R <sup>2</sup> | 0.097                  | 0.326                  |
| Prob > F       | 0.116                  | 0.000                  |

Notes: \*, \*\*, and \*\*\* indicate significance at the 10%, 5% and 1% level, respectively. The omitted category is “Mixedroom/NoInfo”.

different from the “Same”-dummy (F-statistic = 0.08, p-value = 0.7799).

Although it has a negative coefficient, information does not lead to a significantly reduced output level. So the participants in our experiment could not establish tacit collusion. It seems like such collusion is hard to establish in a context of employee compensation. In contrast to that, social interaction significantly decreases the output. So the channel of social interaction seems to have a clearly higher impact than the channel of information. Additionally, information has ambiguous effects. While added information in the Mixed Room treatments has a negative effect on output, it actually has a small positive effect on output in the Same Room treatments. One reason might be the worker’s focus on their output relative to the team member’s output which could be perfectly compared with information. This focus could have introduced a competitive component to the setting, which is rather focused on social interaction when participants just sit together in the office.

### **Result 1 [Output]**

*Only the SN and the SI treatments lead to a significant reduction of output compared to*

## SOCIAL INTERACTION

Table 3.3: Regression of Output Difference on Treatment Dummies and Ability Difference

|                |            |            |
|----------------|------------|------------|
| Info           | -15.833*   | -16.893**  |
|                | (9.217)    | (7.668)    |
| Same           | -19.800*** | -18.086*** |
|                | (6.666)    | (5.837)    |
| Same*Info      | 29.033*    | 27.110**   |
|                | (15.100)   | (11.045)   |
| \Delta Ability |            | 0.135***   |
|                |            | (0.028)    |
| Constant       | 31.500***  | 11.406**   |
|                | (4.636)    | (4.972)    |
| Obs.           | 42         | 42         |
| R <sup>2</sup> | 0.095      | 0.496      |
| Prob> F        | 0.039      | 0.000      |

Notes: \*, \*\*, and \*\*\* indicate significance at the 10%, 5% and 1% level, respectively. The omitted category is “Mixedroom/NoInfo”.

*the MN treatment. The reductions in these treatments are not significantly different from each other.*

### Output Difference

Table 3.3 shows the results of the output difference regression. It regresses output difference on the treatment dummies in the first model and includes the difference of ability in the second model. We focus on the model without the ability difference of two group members. As we only have 42 observations, we want to minimize the number of explanatory variables. The entire model is significant and so is every coefficient. Compared to the omitted category, information as well as same room reduces output difference. But the coefficients are not significantly different from each other (F-statistic = 0.18, p-value = 0.6726). Additionally, all three dummies together are not significantly different from 0 (F-statistic = 0.18, p-value = 0.6726). Additionally, neither the step from the MI treatment (“Information”-dummy) nor the step from the SN treatment (“Same”-dummy) to the SI treatment (all dummies) is significantly different from 0 (F-statistic = 0.46, p-value = 0.5008 and F-statistic = 1.22, p-value = 0.2786, respectively).

Sitting together in one room *or* receiving the output information of the team member has an output difference reducing effect of the same magnitude compared to the baseline (MN). In contrast to that, sitting in one room *and* receiving output information does not reduce output. The reason might again be a reintroduced focus on competition. Due to the low number of observations, we should interpret the results cautiously.

### **Result 2 [Output Difference]**

*The SN and the MI treatments lead to a significant reduction of output difference compared to the MN treatment. But the respective coefficients are not significantly different from each other. The SI treatment does not reduce output difference.*

### **3.5.3 IAB-Data**

Our theoretical analysis and the data suggest that it is suboptimal for the employer to have a relative payment scheme when there is a high degree of social interaction, a fair exchange of productivity information, or both. Were companies implicitly or explicitly knowing this result, we expected relative payments schemes only in work environments with little social interaction and sparse output information among coworkers.

The “Institut für Arbeits- und Berufsforschung”<sup>11</sup> (IAB) in Nuremberg (Germany) maintains the “IAB Betriebspanel”<sup>12</sup>. In this representative panel, about 16,000 establishments in Germany are personally interviewed annually. It includes data points like industry sector, production data, finance data, personnel data, etc. A pilot is conducted half a year in advance of the actual personal interviews which themselves are conducted from June to October of the respective year. In the pilot for the 2012 panel, questions about relative rewards and social interaction were included. Our data is taken from this pilot. A translation of the questionnaire can be found in the appendix.

---

<sup>11</sup>“Institute for Employment Research”

<sup>12</sup>“IAB Establishment Panel” More information can be found on <http://www.iab.de/en/erhebungen/iab-betriebspanel.aspx/>

We could not include our questions in the actual panel and the sample of the pilot is not a random subsample of the panel. So any reasoning based on results of this sample can only be suggestive, not conclusive. We therefore restrain ourselves to a very simple statistical analysis. Statistical tests are based on all valid and non-missing data. However, due to confidentiality reasons, the data may not be presented in tables.

### **3.5.4 Questionnaire Design**

We measured relative incentives through two questions concerning the relative payment and the promotion based on relative performance. For both questions we asked for the existence of such schemes and for the share of employees.

To measure information, we asked whether employees get explicitly informed about the performance of employees with similar tasks. In two further questions, we asked for general social interaction among employees and about the frequency of team events like skiing events or bowling events. Another question dealt with the office composition and therefore asked for both social interaction and in a very broad way.

### **Descriptives**

115 establishments returned valid data for the pilot of the 2012 panel. Their number of employees ranges from 1 to 5,200 with a mean of 376 and a standard deviation of 745.82. Covered industries include trade, construction, services, and production goods. More than half of the establishments are in the sectors of investment/durable goods, trade and the service sector.

### **Results**

As a first step, we build a crosstab of the establishments with(out) relative payment versus the number of establishments with(out) information of workers on the output of others.

## SOCIAL INTERACTION

A marginally significant Chi-squared test supports the impression that information and relative payment seem to be positively related (Pearson  $\text{Chi}^2(1) = 2.7259$ ,  $p = 0.099$ ).

We organize the data of the establishments with(out) relative payment versus the degree of social interaction and we conduct a Chi-squared test. It fails to be significant (Pearson  $\text{chi}^2(4) = 7.3996$ ,  $p = 0.116$ ). But we can find a spearman's rho of -0.1812 at a significance level of 0.0727. So there seems to be a slight negative correlation between the intensity of social interaction and the number of establishments with relative payment. We do not find similar results for the frequency of social interaction and the number of establishments with relative promotion.

However, the results get sharpened if we build a binary index of relative payment and relative promotion. The binary index is set to 1 if either relative payment or relative promotion are implemented in an establishment and 0 otherwise. Such an index makes sense as it eliminates potential ambiguity from special contracts which deal with promotion and relative payment in a special way. Important for our analysis is the general existence of relative rewards in a broad sense. A Chi-squared test (Pearson  $\text{chi}^2(4) = 9.8246$ ,  $p = 0.043$ ) and a spearman rank correlation rho of -0.2302 ( $p = 0.0219$ ) support the notion of negatively correlated frequency of relative rewards and the frequency of social interaction.

In general, the results point towards a negative correlation of social interaction and relative rewards. This is in accordance with our model and our experimental results. In order to maximize output, companies should minimize relative payment if they have a high degree of social interaction and vice versa. On the other hand, we find a marginally significant positive correlation of information and relative payment. This is contrary to our experimental data. We find similar effects of information as social interaction to the output level and the output difference. However, the effect of information in our experiment is not as large as the effect of social interaction. It is not even significant for the output level. So this unclear effect may be the reason why we do not find a negative correlation of information and relative rewards in the IAB data.

### 3.6 Conclusions

We conducted an experiment in order to find out which factors mainly drove the results of the fruit picker experiments in Bandiera et al. (2005). In a relative payment scheme, workers reduced their output level compared to a piece rate payment scheme. However, this was only the case when picking a kind of fruit which allowed the workers to monitor each other and interact socially. It was not clear whether information on the team member's output or social interaction among team members drove this result. In our experiment we find that social interaction reduces the effort strongly compared to no social interaction and no information. Information has a similar effect although not of the same magnitude and not significantly. Information as well as social interaction reduced the output difference of the workers in one team significantly and to the same extent. But information and social interaction did not.

Our results speak in favor of designing work environments in a way that also regards the nature of the respective payment scheme. Relative payments seem to work best in the dimension of output if workers of one group cannot interact socially. So a certain work environment (e.g. open office spaces) may naturally demand a certain payment scheme. Vice versa it is possible that unchangeable payment schemes influence the work environment.

Data taken from a questionnaire sent to more than 100 German establishments support our findings. Relative rewards are uncommon in establishments with a high degree of social interaction. However, relative rewards still exist in work environments with performance information of fellow workers.

One has to bear in mind that the experimental results are only derived from short term contracts for workers, who did not know each other. In Bandiera et al. (2005) the effect was especially strong for befriended workers. So we believe that social ties would strengthen our effects as well.



## 3.7 Appendix

### 3.7.1 Input Mask

Figure 3.2: Input Mask

| Eingegebene Bücher: 1   | Vom Gruppenmitglied eingegebene Bücher: 0 |
|---|---|
| Titel und Untertitel:   | Ökonomische Theorie des Tourismus         |
| Autor (Nachname, Vorname):  | Böventer, Edwin von                       |
| weitere Autoren (falls vorhanden):  | Vahrenkamp, Kai                           |
| Verleger:   | Campus                                    |
| ISBN-Nummer (ohne "-"):   | 3593341115                                |
| Erscheinungsjahr:   | 1989                                      |
| <input type="button" value="Speichern"/> <input type="button" value="Löschen"/> |   |

Notes: Going downwards and from left to right, the fields say “Logged Books”, “Logged Books of Your Team Member” (only displayed in treatments with information), “Title and Subtitle”, “Author (Last Name, First Name)”, “Further Authors (if existent)”, “Publisher”, “ISBN-Number (without “-”)”, “Publication Year”.

### 3.7.2 Payment Handout

### 3.7.3 IAB Questionnaire

Figure 3.3: IAB Questionnaire, English Version, page 1

**R01. a) Are there employees in your company whose salary depends on their relative performance data (i.e. on their work performance in comparison to the work performance of other employees in the same company)?**

Yes .....  No .....  *go to question R02!*

**b) What proportion of employees receives a salary which depends on their relative performance data?**

**If precise data is not available please estimate.**

Percentage of employees:  %

**R02. a) Are there employees in your company whose promotion opportunities depend on their relative performance data (i.e. on their performance in comparison to the performance other employees in the same company)?**

Yes .....  No .....  *go to question R03!*

**b) What proportion of employees' promotion opportunities depend on their relative performance data? If precise data is not available please estimate.**

Proportion of employees:  %

**R03. a) Are there employees in your company who are informed about the work performance of colleagues / others in similar positions?**

Yes .....  No .....  *go to question R04!*

**b) What proportion of employees is informed about the work performance of colleagues in similar positions? If precise data is not available please estimate.**

Percentage of employees:  %

**R04. How intensely do your employees interact socially, i.e. how often do they see each other and communicate with each other?**

Never Very often

—  —  —  —

0 1 2 3 4

SOCIAL INTERACTION

Figure 3.4: IAB Questionnaire, English Version, page 2

**R05. a) What type of office structure is predominant in your company?**

One person (single/individual) office.....

Two person office.....

Group or team office.....

Open plan office.....

Others.....

**b) Are there shared communal multi-function zones for informal communication (e.g. coffee breaks) in your company?**

Yes .....  No .....

**R06. How often do events for your employees (e.g. Christmas parties, ski trips, bowling nights etc.) take place in your company? Please rate using this scale!**

Never  0 — 1  — 2  — 3  — 4  Very often

# Bibliography

- Adams, Christopher**, “Learning in Prediction Markets,” 2006. FTC Working Paper.
- Allais, M.**, “Le comportement de l’homme rationnel devant le risque: Critique des postulats et axiomes de l’école Américaine,” *Econometrica: Journal of the Econometric Society*, 1953, *21*, 503–546.
- Anderhub, V., R. Müller, and C. Schmidt**, “Design and Evaluation of an Economic Experiment via the Internet,” *Journal of Economic Behavior & Organization*, 2001, *46* (2), 227–247.
- Andersen, S., G.W. Harrison, M.I. Lau, and E.E. Rutström**, “Elicitation Using Multiple Price List Formats,” *Experimental Economics*, 2006, *9* (4), 383–405.
- , **J. Fountain, G.W. Harrison, and E.E. Rutström**, “Estimating Subjective Probabilities,” 2010. Copenhagen Business School Working Paper.
- Armantier, O.**, “Do Wealth Differences Affect Fairness Considerations?,” *International Economic Review*, 2006, *47* (2), 391–429.
- Arrow, Kenneth J., Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D. Nelson, George R. Neumann, Marco Ottaviani, Thomas C. Schelling, Robert J. Shiller, Vernon L. Smith, Erik Snowberg, Cass R. Sunstein, Paul C. Tetlock, Philip E. Tetlock, Hal R. Varian, Justin Wolfers, and Eric Zitzewitz**, “The Promise of Prediction Markets,” *Science*, 2008, *320* (5878), 877–878.

## BIBLIOGRAPHY

- Bajari, P. and Ali Hortaçsu**, “Economic Insights from Internet Auctions,” *Journal of Economic Literature*, 2004, *42*, 457–486.
- Bajari, Patrick and Ali Hortaçsu**, “Are Structural Estimates of Auction Models Reasonable? Evidence from Experimental Data,” *Journal of Political Economy*, 2005, *113* (4), 703–741.
- Baltussen, G., G.T. Post, M.J. Van den Assem, and P.P. Wakker**, “Random Incentive Systems in a Dynamic Choice Experiment,” *Experimental Economics*, 2012, *15* (3), 418–443.
- Bandiera, O., I. Barankay, and I. Rasul**, “Social Preferences and the Response to Incentives: Evidence from Personnel Data,” *The Quarterly Journal of Economics*, 2005, *120* (3), 917–962.
- Beattie, J. and Graham Loomes**, “The Impact of Incentives Upon Risky Choice Experiments,” *Journal of Risk and Uncertainty*, 1997, *14*, 149–162.
- Berg, J., R. Forsythe, and T. Rietz**, “What Makes Markets Predict Well? Evidence from the Iowa Electronic Markets,” in “Understanding strategic interaction: Essays in honor of Reinhard Selten,” Springer, 1996, pp. 444–463.
- , —, **F. Nelson, and T. Rietz**, “Results from a Dozen Years of Election Futures Markets Research,” *Handbook of Experimental Economics Results*, 2008, *1*, 742–751.
- Binswanger, Hans P.**, “Attitudes toward Risk: Experimental Measurement in Rural India,” *American Journal of Agricultural Economics*, 1980, *62* (3), 395–407.
- Bossaerts, P., P. Ghirardato, S. Guarnaschelli, and W. Zame**, “Prices and Allocations in Asset Markets with Heterogeneous Attitudes Towards Ambiguity,” *Review of Financial Studies*, 2007, *23*, 1325–1359.
- Brier, G.W.**, “Verification of Forecasts Expressed in Terms of Probability,” *Monthly weather review*, 1950, *78* (1), 1–3.

## BIBLIOGRAPHY

- Bruner, D., M. McKee, and R. Santore**, “Hand in the Cookie Jar: An Experimental Investigation of Equity-Based Compensation and Managerial Fraud,” *Southern Economic Journal*, 2008, 75 (1), 261–278.
- Camerer, C.F.**, “An Experimental Test of Several Generalized Utility Theories,” *Journal of Risk and Uncertainty*, 1989, 2 (1), 61–104.
- Camerer, Colin F. and Teck-Hua Ho**, “Violations of the Betweenness Axiom and Nonlinearity in Probability,” *Journal of Risk & Uncertainty*, 1994, 8 (2), 167 – 196.
- Cason, Timothy N.**, “The Impact of Information Sharing Opportunities on Market Outcomes: An Experimental Study,” *Southern Economic Journal*, 1994, 61 (1), 18–39.
- **and Charles F. Mason**, “Information Sharing and Tacit Collusion in Laboratory Duopoly Markets,” *Economic Inquiry*, 1999, 37 (2), 258–281.
- Charness, G. and U. Gneezy**, “Portfolio Choice and Risk Attitudes: An Experiment,” *Economic Inquiry*, 2010, 48 (1), 133–146.
- Chen, K.Y. and C.R. Plott**, “Information Aggregation Mechanisms: Concept, Design and Implementation for a Sales Forecasting Problem,” 2002. HP Working Paper.
- Cohen, M., J.Y. Jaffray, and T. Said**, “Experimental Comparison of Individual Behavior under Risk and under Uncertainty for Gains and for Losses,” *Organizational Behavior and Human Decision Processes*, 1987, 39 (1), 1–22.
- Cooper, D.J. and K.U. Kühn**, “Communication, Renegotiation, and the Scope for Collusion,” 2011. Florida State University Working Paper.
- Cubitt, Robin P., Chris Starmer, and Robert Sugden**, “On the Validity of the Random Lottery Incentive System,” *Experimental Economics*, 1998, 1, 115–131.
- Curley, S.P., J.F. Yates, and R.A. Abrams**, “Psychological Sources of Ambiguity Avoidance,” *Organizational behavior and human decision processes*, 1986, 38 (2), 230–256.

## BIBLIOGRAPHY

- Eckel, Catherine C. and Philip J. Grossman**, “Men, Women, and Risk Aversion: Experimental Evidence,” in Charles R. Plott and Vernon L. Smith, eds., *Handbook of Experimental Economics Results, Volume 1*, North Holland, 2008, pp. 1061–1076.
- Ellison, Sara Fisher and Catherine Wolfram**, “Coordinating on Lower Prices: Pharmaceutical Pricing under Political Pressure,” *The RAND Journal of Economics*, 2006, *37* (2), 324–340.
- Fama, E.F.**, “The Behavior of Stock-Market Prices,” *Journal of business*, 1965, *38*, 34–105.
- , “Random Walks in Stock Market Prices,” *Financial Analysts Journal*, 1965, *21*, 55–59.
- , “Efficient Capital Markets: A Review of Theory and Empirical Work,” *The Journal of Finance*, 1970, *25* (2), 383–417.
- , “Efficient Capital Markets: II,” *Journal of Finance*, 1991, *46* (5), 1575–1617.
- , **L. Fisher, M.C. Jensen, and R. Roll**, “The Adjustment of Stock Prices to New Information,” *International Economic Review*, 1969, *10* (1), 1–21.
- Farmer, J.D. and A.W. Lo**, “Frontiers of Finance: Evolution and Efficient Markets,” *Proceedings of the National Academy of Sciences*, 1999, *96* (18), 9991–9992.
- Farquhar, P.H.**, “Utility Assessment Methods,” *Management science*, 1984, *30*, 1283–1300.
- Forsythe, R., F. Nelson, G.R. Neumann, and J. Wright**, “Anatomy of an Experimental Political Stock Market,” *The American Economic Review*, 1992, *82*, 1142–1161.
- , **T.A. Rietz, and T.W. Ross**, “Wishes, Expectations and Actions: A Survey on Price Formation in Election Stock Markets,” *Journal of Economic Behavior & Organization*, 1999, *39* (1), 83–110.

## BIBLIOGRAPHY

- Fountain, J. and G.W. Harrison**, “What Do Prediction Markets Predict?,” *Applied Economics Letters*, 2011, 18 (3), 267–272.
- Gjerstad, S.**, “Risk Aversion, Beliefs, and Prediction Market Equilibrium,” 2005. Chapman University Working Paper.
- Goeree, Jacob K., Charles A. Holt, and Thomas R. Palfrey**, “Quantal Response Equilibrium and Overbidding in Private-Value Auctions,” *Journal of Economic Theory*, 2002, 104 (1), 247 – 272.
- , – , and – , “Risk Averse Behavior in Generalized Matching Pennies Games,” *Games and Economic Behavior*, 2003, 45 (1), 97 – 113.
- Greiner, Ben**, “An Online Recruitment System for Economic Experiments,” in Kurt Kremer and Volker Macho, eds., *Forschung und wissenschaftliches Rechnen 2003, GWDG Bericht 63*, Göttingen, Germany: Ges. für Wiss. Datenverarbeitung, 2004, pp. 79–93.
- Grossman, S.**, “On the Efficiency of Competitive Stock Markets Where Traders Have Diverse Information,” *The Journal of Finance*, 1976, 31 (2), 573–585.
- Grossman, S.J. and J.E. Stiglitz**, “On the Impossibility of Informationally Efficient Markets,” *The American Economic Review*, 1980, 70 (3), 393–408.
- Hansen, Lars Peter and Kenneth J. Singleton**, “Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models,” *Econometrica*, 1982, 50 (5), 1269–1286.
- Harbring, C.**, “The Effect of Communication in Incentive Systems - An Experimental Study,” *Managerial and Decision Economics*, 2006, 27 (5), 333–353.
- Harrison, Glenn W.**, “Expected Utility Theory and the Experimentalists,” *Empirical Economics*, 1994, 19, 223–253.



## BIBLIOGRAPHY

- , **Eric Johnson**, **Melayne M. McInnes**, and **E. Elisabeth Rutström**, “Risk Aversion and Incentive Effects: Comment,” *The American Economic Review*, June 2005, *95* (3), 897–901.
- Harrison, G.W.**, “Maximum Likelihood Estimation of Utility Functions Using Stata,” 2008. University of Central Florida Working Paper.
- and **E.E. Rutström**, “Expected Utility Theory and Prospect Theory: One Wedding and a Decent Funeral,” *Experimental Economics*, 2009, *12* (2), 133–158.
- and **J.A. List**, “Field Experiments,” *Journal of Economic Literature*, 2004, *42* (4), 1009–1055.
- , – , and **C. Towe**, “Naturally Occurring Preferences and Exogenous Laboratory Experiments: A Case Study of Risk Aversion,” *Econometrica*, 2007, *75*, 433–458.
- , **M.I. Lau**, and **E.E. Rutström**, “Estimating Risk Attitudes in Denmark: A Field Experiment,” *Scandinavian Journal of Economics*, 2007, *109* (2), 341–368.
- , – , and **M.B. Williams**, “Estimating Individual Discount Rates in Denmark: A Field Experiment,” *The American Economic Review*, 2002, *92* (5), 1606–1617.
- , **S.J. Humphrey**, and **A. Verschoor**, “Choice under Uncertainty: Evidence from Ethiopia, India and Uganda,” *The Economic Journal*, 2010, *120* (543), 80–104.
- Hayek, F.A.**, “The Use of Knowledge in Society,” *The American Economic Review*, 1945, *35*, 519–530.
- Hey, J.D. and C. Orme**, “Investigating Generalizations of Expected Utility Theory Using Experimental Data,” *Econometrica: Journal of the Econometric Society*, 1994, *62*, 1291–1326.
- and **J. Lee**, “Do Subjects Remember the Past?,” *Applied Economics*, 2005, *37* (1), 9–18.

## BIBLIOGRAPHY

- **and** –, “Do Subjects Separate (or Are They Sophisticated)?,” *Experimental Economics*, 2005, 8 (3), 233–265.
- Holt, Charles A.**, “Preference Reversals and the Independence Axiom,” *The American Economic Review*, June 1986, 76 (3), 508–515.
- **and Susan K. Laury**, “Risk Aversion and Incentive Effects,” *The American Economic Review*, December 2002, 92 (5), 1644–1655.
- **and** –, “Risk Aversion and Incentive Effects: New Data without Order Effects,” *The American Economic Review*, June 2005, 95 (3), 902–904.
- Jacobson, S. and R. Petrie**, “Learning from Mistakes: What Do Inconsistent Choices Over Risk Tell Us?,” *Journal of Risk and Uncertainty*, 2009, 38 (2), 143–158.
- Jensen, M.**, “Some Anomalous Evidence Regarding Market Efficiency,” *Journal of Financial Economics*, 1978, 6 (2/3), 95–101.
- Knittel, Christopher R. and Victor Stango**, “Price Ceilings as Focal Points for Tacit Collusion: Evidence from Credit Cards,” *The American Economic Review*, 2003, 93 (5), 1703–1729.
- Knoeber, C.R. and W.N. Thurman**, “Testing the Theory of Tournaments: An Empirical Analysis of Broiler Production,” *Journal of Labor Economics*, 1994, 12, 155–179.
- Kocher, M.G. and S.T. Trautmann**, “Selection into Auctions for Risky and Ambiguous Prospects,” *Economic Inquiry*, 2013, 51, 882–895.
- Lauriola, M. and I.P. Levin**, “Relating Individual Differences in Attitude toward Ambiguity to Risky Choices,” *Journal of Behavioral Decision Making*, 2001, 14 (2), 107–122.
- , – , **and S.S. Hart**, “Common and Distinct Factors in Decision Making under Ambiguity and Risk: A Psychometric Study of Individual Differences,” *Organizational Behavior and Human Decision Processes*, 2007, 104 (2), 130–149.

## BIBLIOGRAPHY

- Laury, Susan K.**, “Pay One or Pay All: Random Selection of One Choice for Payment,” 2005. Working paper 06-13, Andrew Young School of Policy Studies.
- Lazear, E.P.**, “Pay Equality and Industrial Politics,” *Journal of political economy*, 1989, *97*, 561–580.
- **and K.L. Shaw**, “Personnel Economics: The Economist’s View of Human Resources,” *The Journal of Economic Perspectives*, 2007, *21* (4), 91–114.
- **and S. Rosen**, “Rank-Order Tournaments as Optimum Labor Contracts,” *Journal of Political Economy*, 1981, *89*, 841–864.
- List, John A. and Imran Rasul**, “Chapter 2 - Field Experiments in Labor Economics,” in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 4, Part A of *Handbook of Labor Economics*, Elsevier, 2011, pp. 103 – 228.
- Machina, Mark J. and David Schmeidler**, “A More Robust Definition of Subjective Probability,” *Econometrica*, 1992, *60* (4), 745–780.
- Maier, Johannes and Maximilian Ruger**, “Measuring Risk Aversion Model-Independently,” 2011. University of Munich Working Paper.
- Malkiel, B.G.**, “The Efficient Market Hypothesis and its Critics,” *Journal of Economic Perspectives*, 2003, *17*, 59–82.
- Mandelbrot, B.**, “Forecasts of Future Prices, Unbiased Markets, and MartingaleMM-odels,” *Journal of Business*, 1966, *39*, 242–255.
- Manski, C.F.**, “Interpreting the Predictions of Prediction Markets,” 2006. Northwestern University Working Paper.
- Marans, Robert W. and Kent F. Spreckelmeyer**, “Evaluating Open and Conventional Office Design,” *Environment and Behavior*, 1982, *14* (3), 333–351.

## BIBLIOGRAPHY

- Nelson, R.G. and D.A. Bessler**, “Subjective Probabilities and Scoring Rules: Experimental Evidence,” *American Journal of Agricultural Economics*, 1989, 71 (2), 363–369.
- Offerman, T., J. Sonnemans, G. Van De Kuilen, and P.P. Wakker**, “A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes,” *The Review of Economic Studies*, 2009, 76 (4), 1461–1489.
- Oliven, K. and T.A. Rietz**, “Suckers Are Born but Markets Are Made: Individual Rationality, Arbitrage, and Market Efficiency on an Electronic Futures Market,” *Management Science*, 2004, 50 (3), 336–351.
- Ottaviani, M. and P. Sørensen**, “Aggregation of Information and Beliefs: Asset Pricing Lessons from Prediction Markets,” 2012. University of Copenhagen Working Paper.
- Potamites, E. and B. Zhang**, “Measuring Ambiguity Attitudes: A Field Experiment among Small-Scale Stock Investors in China,” 2007. New York University Working Paper.
- Prasad, K. and T.C. Salmon**, “Self Selection and Market Power in Risk Sharing Contracts,” 2010. University of Maryland Working Paper.
- Prelec, D.**, “The Probability Weighting Function,” *Econometrica*, 1998, 66, 497–527.
- Prendergast, C.**, “The Provision of Incentives in Firms,” *Journal of economic literature*, 1999, 37 (1), 7–63.
- Quiggin, J.**, “A Theory of Anticipated Utility,” *Journal of Economic Behavior & Organization*, 1982, 3 (4), 323–343.
- Roll, R.**, “Orange Juice and Weather,” *The American Economic Review*, 1984, 74 (5), 861–880.
- Roth, A.E.**, “Laboratory Experimentation in Economics,” *Economics and Philosophy*, 1986, 2 (02), 245–273.

## BIBLIOGRAPHY

- Rubin, Donald B.**, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, 1987.
- Samuelson, P.A.**, “Proof That Properly Anticipated Prices Fluctuate Randomly,” *The Bell Journal of Economics and Management Science*, 1973, 4 (2), 369–374.
- Samuelson, Paul A. and William D. Nordhaus**, *Economics*, Irwin McGraw-Hill, 1985.
- Schmidt, C. and A. Werwatz**, *How Accurate Do Markets Predict the Outcome of an Event?: The Euro 2000 Soccer Championships Experiment*, Max-Planck-Inst. for Research into Economic Systems, Strategic Interaction Group, 2002.
- Selten, Reinhard**, “Die Strategiemethode zur Erforschung eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes,” in H. Sauermann, ed., *Beiträge zur Experimentellen Wirtschaftsforschung*, Tübingen: Mohr, 1967, pp. 136–168.
- Servan-Schreiber, E., J. Wolfers, D.M. Pennock, and B. Galebach**, “Prediction Markets: Does Money Matter?,” *Electronic Markets*, 2004, 14 (3), 243–251.
- Slamka, C., A. Soukhoroukova, and M. Spann**, “Event Studies in Real -and Play-Money Prediction Markets,” *The Journal of Prediction Markets*, 2008, 2 (2), 53–70.
- Smith, V.L.**, “Markets as Economizers of Information: Experimental Examination of the ‘Hayek Hypothesis’,” *Economic Inquiry*, 1982, 20 (2), 165–179.
- Snowberg, E., J. Wolfers, and E. Zitzewitz**, “Prediction Markets for Economic Forecasting,” 2012. California Institute of Technology Working Paper.
- Stahl, D.O. and E. Haruvy**, “Other-Regarding Preferences: Egalitarian Warm Glow, Empathy, and Group Size,” *Journal of Economic Behavior & Organization*, 2006, 61 (1), 20–41.

## BIBLIOGRAPHY

- Starmer, C. and R. Sugden**, “Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation,” *The American Economic Review*, 1991, *81*, 971–978.
- Sutter, M. and C. Strassmair**, “Communication, Cooperation and Collusion in Team Tournaments – An Experimental Study,” *Games and Economic Behavior*, 2009, *66* (1), 506–525.
- Tirole, J.**, “On the Possibility of Speculation under Rational Expectations,” *Econometrica*, 1982, *50* (5), 1163–1181.
- Wilcox, Nathaniel T.**, “Lottery Choice: Incentives, Complexity and Decision Time,” *Economic Journal*, 1993, *103*, 1397–1417.
- Wolfers, J. and E. Zitzewitz**, “Prediction Markets,” *Journal of Economic Perspectives*, 2004, *18*, 107–126.
- **and** –, “Interpreting Prediction Market Prices as Probabilities,” 2007. National Bureau Of Economic Research Massachusetts.
- Zahn, G. Lawrence**, “Face-to-Face Communication in an Office Setting: The Effects of Position, Proximity, and Exposure,” *Communication Research*, 1991, *18* (6), 737–754.