Prediction Markets versus Alternative Methods

Empirical Tests of Accuracy and Acceptability

Zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der Fakultät für
Wirtschaftswissenschaften
der Universität Karlsruhe(TH)

genehmigte
DISSERTATION

von

Dipl.-Volkswirt, Dipl.-Wirtsch.Inf. Andreas Graefe

Tag der mündlichen Prüfung:	25. Mai 2009
Referent:	Prof. Dr. Christof Weinhardt
Korreferent:	Prof. Dr. Armin Grunwald

2009, Karlsruhe

Acknowledgments

I am indebted to many people for their support and collaboration throughout this Ph.D. project. I would like to thank my supervisor Christof Weinhardt for his strong support and for continuously providing valuable insight and comments. I would also like to thank Armin Grunwald for co-advising this thesis and giving me the opportunity to work in the highly interdisciplinary research environment at the Institute for Technology Assessment and Systems Analysis (ITAS) at the Karlsruhe Institute of Technology. There, I was able to discuss and exchange ideas with colleagues from different fields and was involved in research projects that were beyond the focus of my thesis, which provided a stimulating environment to grow and learn. Thereby, special thanks to my friend and mentor Carsten Orwat who navigated me through my first steps in the academic world and was always there with helpful advice. I would also like to express my gratitude to my colleagues at the Information & Market Engineering Group at the University of Karlsruhe for providing constructive feedback and making it a pleasure to attend the numerous doctoral seminars. Special thanks to Stefan Luckner for proofreading the manuscript and providing helpful hints and pointers.

I would especially like to thank J. Scott Armstrong for providing the opportunity to work with him for the past two years as a visiting scholar at the University of Pennsylvania's Wharton School. Scott has been an inspiring mentor and a wonderful friend. Under his guidance I was constantly pushed and challenged. He made my stay an exciting and fruitful

experience and helped me discover my passion for research. Moreover, his and his family's warm welcome made me fall in love with the city and people of Philadelphia. Thanks, Scott.

I dedicate this thesis to my family, my parents Lotte and Günther and my sister Kerstin. I have relied on their love and encouragement throughout my life. Finally, I would like to thank Jamie for her love, patience, and emotional support – and for reminding me of the life beyond research.

Abstract

The success of prediction markets in the field of election forecasting made them increasingly appealing to organizations and a number of companies started to experiment with them. However, despite widespread initial interest and years of experimental use, there are no major organizations known to use prediction markets as an integral part of their forecasting activities. Prediction markets have not become an established forecasting method yet.

The reasons for this are manifold. Aside from election and sports forecasting, the number of empirical studies that analyze prediction markets' performance is limited. In addition, the studies are often small scale or compare the method to weak benchmarks. Since the emergence of the field, no meta-analysis has been published to analyze prediction markets' accuracy. Furthermore, practical experience indicates that cognitive and organizational barriers thwart the implementation of prediction markets within organizations.

This work provides further empirical evidence on the performance of prediction markets and analyzes the method's acceptability among participants. Results from a field experiment showed that prediction markets performed equally well as the Delphi method for long-term forecasting problems. Similar results were derived from a laboratory experiment on a quantitative judgment task: overall, prediction markets performed equally to Delphi, nominal groups, and meetings. Furthermore, they appeared to be particularly valuable for problems where multiple group members had valid information.

However, laboratory experiment participants had comparably unfavorable perceptions of prediction markets, particularly in terms of difficulty of participation, which may lead to low confidence in market results. This seemed to be confirmed by the fact that market results were discounted more often than results from the three benchmark approaches. Yet, in discounting the market results, participants did not improve accuracy; they harmed it. The results suggested that prediction market participants were unable to judge the quality of market results and, thus, should have refrained from revising them.

The empirical evidence from this work supports the value of prediction markets for forecasting. However, it also revealed that prediction markets are afflicted with unfavorable perceptions of participants. This conformed to practical experience indicating barriers to the method's implementation within organizations. Future research aimed at identifying and overcoming these barriers is of crucial importance. There is a need for further empirical studies that analyze performance of prediction markets for different types of problems and in different settings. This should involve the use of prediction markets in conjunction with traditional means of forecasting. In addition, market engineering should search for ways to make market platforms more accessible, particularly to non-experience participants.

Contents

Figures

Tables

Abbreviations

APE	Absolute percentage error
CDA	Continuous double auction
DARPA	Defense Advanced Research Project Agency
EPIS	European Perspectives on Information Society
FX	Foresight Exchange
FTF	Face-to-face meeting
HSX	Hollywood Stock Exchange
ICT	Information and Communication Technologies
IEM	Iowa Electronic Markets
IRB	Institutional Review Board
JAS	Judge-advisor system
MAPE	Mean absolute percentage error
MdAPE	Median absolute percentage error
MOV	Mean order volume
NGT	Nominal group technique
PAM	Policy Analysis Market
PM	Prediction Markets
RT Delphi	Real-time Delphi
TechForX	Technology Foresight Exchange

Chapter 1

Introduction

Prediction markets were quite popular already in the late 19th century. In analyzing historical markets that were operated for betting on the 15 U.S. presidential elections from 1884 and 1940, Rhode and Strumpf (2004, p.127) found that these markets "did a remarkable job forecasting elections in an era before scientific polling". At certain times, trading activity in these markets was higher than in the stock exchanges on Wall Street, and, during some election campaigns, newspapers like the *New York Times* reported market prices on a nearly daily basis. Nonetheless, with increasing availability of other forms of gambling (like horse races) and the rise of opinion polls – which were not subject to moral concerns associated with gambling – presidential betting markets disappeared after 1940.

In recent years, there has been a resurgence of interest in prediction markets in the field of forecasting. In 1988, the *Iowa Electronic Market* (IEM)[1] was launched to predict the

[1] In 1988, the Henry B. Tippie College of Business at the University of Iowa launched the Iowa Electronic Markets (IEM), the first prediction market that facilitated participation over the internet (http://www.biz.uiowa.edu/iem/). With its first attempt to predict the outcome of the 1988 U.S. presidential election, the IEM provided more accurate forecasts than traditional opinion polls. And it did so for all consecutive elections. In analyzing 964 polls for the five presidential elections from 1988 to 2004,

outcome of the U.S. presidential elections of the same year. Its initial success, accompanied by the rise of the internet, ignited the interest of researchers. Since the mid-1990s, various studies have been published that demonstrated accuracy of prediction markets in areas beyond election forecasting, for example, for predicting sports events or business figures. In response to increasing academic interest, the *Journal of Prediction Markets* was initiated in 2007. In 2008, in an effort to transfer research findings to scholars and practitioners in the field, the *Special Interest Group on Prediction Markets*[2] was launched on behalf of the *International Institute of Forecasters*.[3]

The final boost in popularity for prediction markets can be traced back to two events. Ironically, the cancellation of the Policy Analysis Market (PAM)[4] in 2003, which was covered by more than 600 media articles, initially made a broad public aware of prediction

---

Berg et al. (2008a) found that the respective market forecasts were closer to the actual election results 74% of the time. This superior performance compared to individual polls was replicated for the 2008 election (Berg et al. 2008b). For further information see Section 2.3.1.

[2] http://www.marketsforforecasting.com

[3] http://www.forecasters.org

[4] From 2001 to 2003, the Defense Advanced Research Project Agency (DARPA) of the U.S. government sponsored the FutureMAP project, also known as the Policy Analysis Market (PAM). The original goal of this project was to improve existing intelligence institutions by predicting military and political instability around the world, how the U.S. would affect such instabilities, and vice versa. Later, the focus was narrowed to predict five parameters for each of eight nations in the Middle East: military activity, political instability, economic growth, U.S. military activity, and U.S. financial involvement. In addition, traders should predict additional parameters like U.S. GDP growth, world trade, or total U.S. military casualties.

On July 28, 2003, shortly before the scheduled start of PAM on September 1, two Democratic Senators held a press conference accusing the U.S. Department of Defense to plan a 'terror market' for people to bet on terrorist events. The topic caught the interest of the media. During the next two days, 128 media articles were published and most of them cast PAM in an unfavorable light. Not surprisingly, PAM was rapidly terminated.

Later, Hanson (2007), who was involved in the project, conducted a statistical news analysis on more than 600 media articles that mentioned PAM. He found that the more informed articles favored PAM. Yet, the political decision to dismiss PAM was made and it is unlikely that it will be reversed anytime soon. See Hanson (2007) for a review of the origin and development of the project.

markets. Second, James Surowiecki's bestselling book 'The Wisdom of Crowds'[5], published in 2004, described prediction markets as one of many ways to harness collective intelligence. In the following year, prediction markets were listed on the *Gartner Hype Cycle* (Fenn & Linden 2005) and soon major business consultancies saw them as an emerging trend (Manyika et al. 2007).

Not surprisingly, prediction markets became increasingly appealing to companies. Since 2005, the media regularly reports on companies experimenting with prediction markets. For example, news articles published in *BusinessWeek* (King 2006) or *IEEE Spectrum* (Cherry 2007) named various companies (e.g. Microsoft, Yahoo, Intel, Eli Lilly or Nokia) that have experimented with prediction markets to improve their internal decision-making processes, for example by forecasting the success of new products, commodity prices, or sales figures. With almost 1,500 employees participating between 2005 and 2007 the largest known internal prediction market ran at Google (Cowgill et al. 2008). The Google markets aimed at predicting future demand (e.g. 'number of Gmail users by the end of the quarter'), performance (e.g. 'Google Talk quality rating'), company news (e.g. 'Google's Russia office to open by…'), or industry news (e.g. 'Will Apple release an Intel-based Mac?'). For an overview of ongoing media coverage about the use of prediction markets, see the *Special Interest Group on Prediction Markets*.

The evidence from the literature suggests that prediction markets have potential to improve on forecasting. However, published case studies are few and they often draw on small

[5] In presenting examples of situations in which averaged crowd opinions outperformed single experts, James Surowiecki's 2004 book 'The Wisdom of Crowds' demonstrated the power of collective intelligence to a broad audience. Yet, the title of the book is misleading: crowds are not wise when acting together. The actual conclusion of the book is that the combined knowledge of many individual contains wisdom, when the individuals act independently. Surowiecki described prediction markets as one of many ways to harness crowd opinion and, thus, contributed largely to their awareness level.

samples. In addition, prediction markets have often been compared to weak benchmarks like individual polls, individual expert judgments, or naïve models (like random walk).

To date, it is not known of any major organization that has implemented prediction markets as an integral part of their forecasting activities. Although larger organizations, in particular technology and pharmaceutical companies, were the first to experiment with the approach, its use has not spread to other domains.[6] As a result, mainstream awareness of the approach is still limited. Currently, prediction markets are undergoing the typical life (or hype) cycle of innovative technologies. As framed by Cain and Drakos (2008), prediction markets have overcome the 'peak of inflated expectations' and are entering the 'through of disillusionment': early adopters – scientists as well as practitioners – who had overestimated prediction markets' accuracy and overall usefulness, are now to some extent disenchanted.

1.1 Defining the Scope of the Work

Before describing the goals of this work, this section outlines what lies beyond the scope of this research.

This work does not provide a thorough analysis of questions related to the stream of research referred to as *market engineering* (Weinhardt et al. 2003). Market engineering deals with questions of how to analyze, design, implement, and enhance all types of electronic markets in order to establish conditions for efficient markets. Such questions are important and have received a lot of attention in the field of prediction markets thus far. In particular, several studies have focuses on the design of trading mechanisms and incentive schemes.

[6] In a talk at the *Forecasting Summit* (Graefe 2008d) – a conference for business forecasting practitioners – the audience of approximately 50 people was asked if they were using prediction markets within their companies. Nobody used prediction markets – and only a handful had even heard of them before.

Market inefficiencies due to low liquidity are well-known for prediction markets that are based on the continuous double auction (CDA) mechanism. They have been reported for thin markets in the laboratory (Chen et al. (2004), Rietz (2005)) as well as for large-scale markets like the IEM (Forsythe et al. 1999, Oliven & Rietz 2004) or Tradesports.com (Tetlock 2008), which are populated by self-selected – and often experienced – traders. Even though all authors reported high forecasting accuracy, such inefficiencies are often raised a key concern in using prediction markets. In particular, it is often assumed that low liquidity and trading activity have a negative impact on market accuracy. In response, researchers have developed automated market maker mechanisms to provide additional liquidity (Hanson 2003, Pennock 2004). For an overview and discussion of the several market mechanisms see Luckner (2008). However, the assumption that low liquidity harms forecasting accuracy has recently been questioned by Tetlock (2008). In analyzing three years of trading data from TradeSports.com, he found that more liquid markets were not more accurate. In fact, they were sometimes even less accurate.

Another question that attracted the interest of researchers is the relative performance of play-money and real-money markets. Two studies addressed this question. While Servan-Schreiber et al. (2004) could not identify differences in accuracy between play-money and real-money markets for sports events, Rosenbloom and Notz (2006) found real-money markets to be more accurate for non-sports events. In reviewing both studies, Luckner (2008, p.76) concluded that "the impact of real money vs. play money on the accuracy of predictions is not completely understood and clarified".

For the application of prediction markets within companies, play-money markets appear to be most suitable. This raises questions about the optimal design of incentive schemes in such markets. Using the STOCCER market, which has been launched to predict the outcome of the 2006 FIFA World Cup, Luckner and Weinhardt (2007) conducted a field experiment to compare different incentive schemes for play-money markets. They found

that, for the common assumption that traders are risk-averse, incentive schemes should be designed as rank-order tournaments.

Furthermore, it is not the goal of this work to discuss impacts of trader biases. For example, Wolfers and Zitzewitz (2004) analyzed financial contracts traded at Tradesports.com and identified a favorite longshot-bias. According to this bias, which is well-known in horse races (Thaler & Ziemba 1988), traders overestimate longshots (i.e. extremely unlikely events) and underestimate favorites. Others showed that market participants trade according to their desires. In analyzing trading behavior in the IEM, Forsythe et al. (1992) found that traders were biased by political preferences: they bought more contracts of the candidate they supported than they sold. This 'wishful thinking effect' was replicated by Luckner (2008) for sports predictions. Based on data from the STOCCER FIFA World Cup market, he showed that traders held and bought more contracts of the team that represented their nationality. Yet, despite biased traders in both studies, Forsythe et al. (1992) as well as Luckner (2008) concluded that the markets provided accurate predictions. Apparently, prediction markets do not require all traders to behave rational. According to the marginal trader hypothesis by Forsythe et al. (1992), a few rational traders – who invest more and trade more actively – are sufficient for the market to perform well.

In sum, we know that prediction markets are inefficient and traders are biased. Thus, market engineering will remain an important stream of research. It can aid in the design of systems that deal with inefficiencies of markets and traders' biases or risk attitudes. In turn, this can make prediction markets more attractive, accurate, and, thus, more valuable for organizations.

1.2 Motivation

Despite all inefficiencies and biases, the track record of prediction markets is good and the studies available to date report high accuracy. Yet, it seems that prediction markets are not

used as often as they should for forecasting or aggregating information from people. In approaching prediction markets from a practical viewpoint, this work discusses some of the obstacles that hinder their implementation.

1.2.1 Lack of empirical evidence

There is a remaining need for empirical studies that validate the performance of prediction markets. As will be outlined in Chapter 2, available studies are limited and often small scale. Since the emergence of the field, no meta-analysis has been published that reviewed the literature and analyzed prediction markets' accuracy. In particular, there is need for studies that analyze the *relative* performance of prediction markets. As long as there is no evidence that prediction markets are superior to alternative mechanisms, organizations have no need to depart from their status quo. This work contributes to the literature by comparing prediction market accuracy to traditional approaches of information aggregation and judgmental forecasting like face-to-face meetings, nominal groups, and the Delphi method.

Furthermore, so far, the track record of prediction markets is based on forecasting events in the near future. However, forecasting must not be limited to the near future but has to consider the long-term. Prediction markets might increase their appeal to organizations if they can be shown to be applicable for long-term forecasting.

The literature provides little evidence on whether prediction markets are applicable for long-term problems whose outcome cannot be judged at the time the prediction is being made. Results from an analysis of the *Foresight Exchange (FX)*[7], a prediction market aiming at assessing long-term developments, suggested that prediction markets might perform well for longer forecasting horizons. For 161 contracts that referred to 'yes' or 'no' questions, Pennock et al. (2001a) recorded FX forecasts thirty days before the respective outcome was known. They found that the FX forecasts strongly correlated with outcome frequencies.

[7] http://www.ideosphere.com/

However, they did not compare the results to a benchmark forecast. It is still an open question whether prediction markets are appropriate for long-term forecasting. Further empirical studies are necessary to assess the forecasting performance of markets for longer time horizons.

1.2.2 Cognitive barriers

Practical experience indicates that people have problems in understanding how prediction markets work. In particular, they have problems in translating information into market prices (Green et al. 2007) and do not understand how to read them. For example, in line with the 2008 financial crisis at Wall Street, the real-money prediction market intrade.com launched a contract to predict whether the U.S. government will pass a bail out. With a market price traded at $80, this market forecasted an 80% probability that the bail out will go through. A staff writer at CNNMoney.com – who could be expected to possess a certain understanding of financial instruments – wrongly interpreted the market prices in stating that *80% of the participants thought* that the event will come true (Rooney 2008).

Misunderstanding of how prediction markets work might be critical for their acceptance by participants as well as decision-makers. Indeed, as described in Section 1.1, for the market to provide accurate forecasts it is *not* required that participants completely understand the functioning of prediction markets – or behave fully rational. However, if people do not fully understand how to reveal information, how information is aggregated into market prices, and how to interpret market prices, it is doubtful that they have favorable perceptions of the mechanism and, thus, have confidence in the results. As a result, people might insist on using alternative mechanisms that are easier to understand – even if they might be less accurate.

1.2.3 Organizational barriers

Prediction markets change the way decisions are traditionally made in organizations. This may result in organizational barriers.

When making decisions, people tend to rely on expert advice. Experts are believed to possess superior information and to be able to utilize this information effectively. Thus, they are expected to be superior in prediction, especially in situations involving high uncertainty. Yet, findings from empirical research dispute the value of experts in forecasting. In reviewing the literature, Armstrong (1980) found that there is little benefit to expertise when predicting change. Although expertise beyond a minimum level was found to lead to more accurate forecasts, additional expertise did not improve accuracy. In fact, Armstrong found some evidence that accuracy might even decrease with increasing expertise.

More recent, Tetlock (2006) published results from a longitudinal study in which he analyzed more than 28,000 expert assessments of future political and economic events. He found that experts barely – if at all – outperformed non-experts. In addition, they were only slightly more accurate than simple rules like 'predict no change' or 'predict most recent rate of change'. Tetlock referred to these benchmarks as 'mindless competition'. In addition, the experts were better at making predictions outside their areas of expertise than in them.

The – perhaps counterintuitive – finding that decision or forecasting accuracy can decrease with increasing expertise has been illustrated in a study of Arkes and Harkness (1980), who analyzed the accuracy of clinical decisions. The authors showed speech therapy students a list of symptoms and asked them to make a diagnosis of Down's syndrome. Even though the common symptom 'fissured tongue' has not been presented to the students, they tended to remember having seen it. Non-experts are less likely to make this error, simply because they have less background knowledge.

So why do decision-makers tend to rely on expert advice? In response to this question, Armstrong (1980, p.16) developed the "Seer-sucker Theory", stating that "no matter how much evidence exists that seers do not exist, seers will find suckers". He argued that, despite all evidence that disputes the value of experts in forecasting, expert judgments increase credibility and reliability of decisions – and allow decision-makers for avoiding responsibility.

In being built on the idea of the 'wisdom of crowds' (Surowiecki 2004), prediction markets do not necessarily aim at separating experts from non-experts. Although it is possible to limit who can participate, prediction markets are usually open for everyone. Yet, the idea of involving amateur knowledge into decision-making can be a barrier for prediction markets as has become evident with the dismissal of the Policy Analysis Market (see Footnote 4). People had concerns that, by using prediction markets, skilled professionals would be replaced by unskilled self-chosen amateurs (Hanson 2006).

There might be other situations in which publicly available information on prediction markets is not desirable. Imagine an internal market that was set up to estimate whether a project will be successful and imagine that this market forecasts that the project will fail. Such information might discourage project members. If they get the sense that their project will fail, they might decrease their effort. As a result, the project will fail and the forecast becomes a self-fulfilling prophecy. Or, the project manager might have an interest in keeping such a forecast secret. If his job or future career depends on the success of the project, he has an incentive to conceal – or at least defer – bad information.[8]

Such situations are problematic for forecasting in general and might occur regardless of the applied method. However, the difference is that prediction markets make the forecast

[8] These were concerns raised by business forecasting practitioners in the discussion following a talk at the *Forecasting Summit* (Graefe 2008d).

transparent. In setting up a market to answer a question, one also reveals the answer – at least to the market participants. To stay with the above example: a forecast that predicts a project to fail might call for interventions by the project manager to bring the project back on track. If he is reluctant to perform such interventions – and the project fails indeed – he will have to explain why he ignored information that was obvious to everyone. Thus, prediction markets might be perceived as a danger to existing management structures. They change the process of decision-making and make wrong decisions transparent.

This raises the question about how decision-makers should use market results. Should they completely rely on prediction market outcomes? Or should they consider market results as an input (i.e. as one source of information) when making their final judgment?

1.3 Research Questions

The goal of this work is to address some of the remaining barriers that have been identified to hinder the further implementation of prediction markets. This section outlines the specific research questions dealt with in this work.

Research question 1:

How do prediction markets perform for long-term forecasting?

As mentioned earlier, the literature provides little evidence about the performance of prediction markets for forecasting the long-term future. This is because of two reasons. First, for suchlike problems, it is impossible to define a clear pay-off function, which is a basic requirement for the proper functioning of the market mechanism and provides incentives for participation. Second, it is impossible to validate the accuracy of the results before the actual outcome can be observed. It just cannot be known soon enough.

Thus, the question arises how the performance of prediction markets can be evaluated for suchlike problems. A way to solve this problem is to use established methods as external benchmarks for validating market outcomes. This work presents results from the TechForX field experiment, in which two prediction markets were launched in parallel to a well-established approach in judgmental long-term forecasting: the Delphi method. The goal of this exercise, which was conducted in line with a European foresight project, was to assess long-term technological trends in the creative content industries.

Research question 2:

Do experts have value in prediction markets?

Experience with the PAM has revealed people's concerns with involving unskilled amateurs in the decision-making process by using prediction markets (Hanson 2006). People feared that inexperienced participants might distort market results. In general, these concerns appear to be unwarranted. Even if inexperienced participants would execute uninformed trades, such 'noise traders' provide additional liquidity. According to assumptions of rational models of liquidity provision, this enhances incentives for other participants to get involved and informed – and to reveal their information through trading. In response, it is often assumed that market forecasts might become even more accurate. This is supported by earlier research, which has shown that prediction markets provide accurate forecasts in spite of biased traders (Forsythe et al. 1999, Luckner 2008) or even intentional attempts to manipulate market prices. Most empirical studies to date showed that manipulation has not been successful historically (Rhode & Strumpf 2004), in the laboratory (Hanson et al. 2006), or in the field (Camerer 1998). Only one study reported successful manipulation of prices at the IEM (Hansen et al. 2004). In reviewing studies of price manipulation, Wolfers and Zitzewitz (2004) concluded that, besides a short transition phase, none of the known attacks had a noticeable influence on the prices.

Nonetheless, it is often assumed that experts possess superior knowledge. Although the literature disputes the value of expertise in forecasting change (see Section 1.2.3), the question arises whether expert-based markets can improve forecasting performance. To date, no study has analyzed the relative performance of experts and amateurs in prediction markets. This work addresses this deficit by analyzing additional data from the TechForX field experiment on forecasting long-term technological trends. The outcomes of two markets, one comprised of experts and one of amateurs (students), were compared to the results of a Delphi study. In addition, trading behavior in both markets was analyzed in terms of whether there is support for the assumption that experts possess superior knowledge.

Research question 3:

How do prediction markets perform compared to traditional approaches?

A forecasting problem can be approached in various ways. If sufficient data for mathematical analyses are available, one often uses quantitative (statistical) methods. In assuming that the future will not be substantially different from the past, such approaches aim at predicting the future based on historical data. Thereby, the selection of the appropriate method depends on additional factors like the type of data or whether there is good domain knowledge about the data. However, often the future cannot be predicted from the past and, in some situations, there is not enough data available for quantitative analyses. Furthermore, additional qualitative information can help to increase accuracy – or acceptability – of forecasts. In such situations, one should incorporate human judgment by using evidence-based *judgmental* methods. For guidelines about which method to use in a particular situation, see the *Selection Tree of Forecasting* at ForPrin.com.

A variety of judgmental approaches can help to elicit information from groups and can thus be valuable for forecasting. Most commonly, organizations rely on traditional face-to-face

meetings, even though meetings are expensive and subject to many biases. In reporting on personal experience and findings from the literature, Armstrong (2006) concluded that it is difficult to find reasons to support the use of meetings for forecasting. Instead, evidence suggests that structured approaches like nominal groups or the Delphi method allow for more accurate forecasts than meetings.

In aggregating dispersed information from people, prediction markets are a judgmental forecasting approach and should thus be compared to similar mechanisms. Yet, little is known about the relative performance of prediction markets compared to unstructured meetings or other structured approaches. To date, the author does not know of any study that compared prediction markets to nominal groups or the Delphi method.

This work reports on a laboratory experiment that was conducted to compare relative accuracy of prediction markets, traditional face-to-face meetings, nominal groups and the Delphi method on a quantitative judgment task.

Research question 4:

How do people perceive participation in prediction markets?

Practical experience indicates that people have problems in understanding how prediction markets work. If so, one can assume that people are not satisfied with participation in a market and, therefore, might not trust the results. So far, the author has been unable to find studies that analyzed how people perceive participation in prediction markets and how this affects confidence in the market outcomes.

People's perception of a forecasting or decision-making process can be crucial for the acceptability of its results. If they feel dissatisfied with the process, its outcome may not be adopted – even if highly accurate. (That said, a process that is satisfying for participants –

like meetings often are – may not necessarily lead to accurate results.) Evidence for poor satisfaction with prediction markets would call for further research in the field of market engineering. In particular, it would create needs for market developers to make their systems more accessible and user-friendly, especially for non-experienced participants.

This work examines participants' perceptions of prediction markets by analyzing data from ex-post evaluation questionnaires of the laboratory experiment on the relative performance of prediction markets.

Research question 5

How do people use market results? How should they use them?

If confidence in prediction markets is low, it is unlikely that decision-makers fully rely on market results. Thus, two questions arise: (1) How *do* people make use of market results? Do they fully rely on them or do they revise them when making decisions? If people show a tendency to revise market results, this, in turn, would suggest low confidence. (2) How *should* people use market results? Should they fully rely on the market results or should they use them as a source of information when making final decisions. In other words, does the strategy of revising market results improve accuracy or should people stick with the market results?

These questions will be addressed by analyzing additional data from the laboratory experiment. In particular, prediction markets are examined as a so-called judge-advisor system.

1.4 Overview and Structure

This work is structured in nine chapters. After the present introduction, Chapter 2 explains the concept of prediction markets, summarizes studies from the literature on their performance, and describes promising characteristics of prediction markets to improve on traditional forecasting methods.

Chapter 3 provides an overview of the research methodology. It explains the three benchmark methods to which prediction markets were compared and defines the type of tasks that have been analyzed. In addition, the chapter outlines the experimental settings of the underlying empirical work. In particular, it describes the design of the TechForX field experiment on long-term forecasting as well as the laboratory experiment on the comparison of prediction markets to meetings, nominal groups, and the Delphi method.

In reporting on results from the TechForX field experiment, Chapter 4 analyzes how prediction markets perform for long-term forecasting problems. It will be shown that the outcomes of two prediction markets strongly correlated with the results of an expert-based Delphi study in forecasting technological trends. These results conformed to findings from earlier research in the field of market research. This led to conclude that prediction markets do not have to be limited to events in the near future but appear to provide reliable results also for long-term forecasting problems.

Chapter 5 addresses the question of whether it can be valuable to restrict participation in prediction markets to experts. Two markets from the TechForX field experiment (one expert and one student market) were compared in terms of market outcomes as well as trading behavior of participants. In considering the parallel running Delphi study as a benchmark, the results from both markets did not differ. This conformed to findings from the literature that experts have little value in forecasting change (see Section 1.2.3). However, an analysis of the order volumes – which were interpreted as a measure of

participants' confidence – revealed differences between experts and students. In particular, the results suggested that experts revealed information well-considered whereas students 'just traded'. In revealing such information to decision-makers, prediction markets appear to be able to improve on traditional forecasting approaches.

Chapter 6 analyzes data from a laboratory experiment on the relative performance of prediction markets compared to traditional face-to-face meetings (FTF), nominal groups (NGT), and the Delphi method on a quantitative judgment task. As one would expect, the three structured approaches were more accurate than FTF. Delphi performed best, followed by NGT and prediction markets, although overall differences were small. However, the relative accuracy of FTF and structured methods appeared to be affected by the difficulty of questions. While meetings tended to be superior for hard-to-estimate questions, they were outperformed by the three structured approaches for easier questions.

In analyzing an ex-post evaluation questionnaire from this laboratory experiment, Chapter 7 analyzed participants' perceptions of prediction markets. Besides confidence in the results, participants were asked to reveal how they perceived the group as well as the group process as a whole. The results were compared to participants' perceptions of meetings, the Delphi method, and nominal groups. While participants favored methods involving personal communication like NGT and FTF, prediction markets were rated least favorable on most categories.

In reporting on additional data from the laboratory experiment, Chapter 8 examines how people made use of prediction market results – and how they should have made use of them. After participating in the group process and being presented with its results, participants were asked to provide their final individual estimates. The results conformed to basic findings from the literature on advice-taking. In particular, participants tended to revise the group results, although relying on them would have improved decision accuracy. Participants who were more confident in their individual estimate than in the group result

discounted the group result more heavily. In addition, the results revealed that participants in prediction markets (as well as in Delphi) appeared to fail in judging the quality of the group results. Furthermore, prediction market participants revised the group results most often and, thereby, harmed decision accuracy.

Chapter 9 summarizes the findings and discusses the contributions of this work. In addition, it raises questions that appear to be promising for future research.

Being common practice, the data presented in the Chapters 4 to 8 was analyzed using tests of statistical significance.[9] Following the recommendations of Armstrong (2007), it is suggested that readers focus on effect sizes in the data. In being reported in footnotes, readers can skip tests of statistical significance if they like. Significance levels are indicated in tables using asterisks.[10] For full disclosure, all data that has been analyzed in this study can be found online.[11] For the calculation of the error measures used in this work, see the Technical Appendix T1.

1.5 Related Presentations and Publications

Most of this work has been presented to – and discussed with – fellow scholars at international scientific conferences and parts of it have been published in various scientific journals.

The design for the TechForX field experiment, which provided the basis for the analyses in Chapters 4 and 5, was presented at the *27th International Symposium on Forecasting* in New

[9] All statistics were calculated using SPSS 15.0.

[10] The following coding was used to indicate levels of significance: * ($p \leq 0.05$) and ** ($p \leq 0.01$).

[11] The links to the data are provided in the respective chapters or can be found in the supporting online material in the Appendix.

York City (Graefe 2007a). Results from the laboratory experiment, reported in Chapters 6 to 8, were presented at the *28th International Symposium on Forecasting* in Nice, France (Graefe 2008b), the *Third Workshop on Prediction Markets* in Chicago (Graefe 2008a, Graefe 2008c), the *2008 INFORMS Annual Meeting* in Washington D.C. (Graefe 2008f), and in an invited talk given at the *Forecasting Summit* in Boston (Graefe 2008d). Furthermore, drafts and ideas of this research were presented at the *Third International Conference on Organizational Foresight* in Glasgow, UK (Graefe 2007e), the *Third International Conference on Technology, Knowledge, and Society* in Cambridge, UK (Graefe 2007d), and the *Second Conference of the TA Network* in Berlin (Graefe 2006).

Parts of Chapters 4 and 5 have been published in the *Journal of Prediction Markets* (Graefe & Weinhardt 2008). The results described in Chapters 6 and 7 are under review with the *International Journal of Forecasting* (Graefe & Armstrong 2008b). Other research papers related to this work have been published – or accepted for publication – in various academic journals including *Futures* (Graefe et al. 2009c) or *Foresight* (Graefe 2008e, Graefe et al. 2009a, Green et al. 2007). Further relevant publications or working papers include Graefe (2007b, 2007c), Graefe and Orwat (2007), Graefe and Armstrong (2008a) and Graefe et al. (2009b).

Chapter 2

Prediction Markets

For most problems in our society, it cannot be expected that a single individual has all the necessary information to solve them. Rather, the relevant bits of information are most likely dispersed among many people. If it is possible to aggregate and utilize this dispersed knowledge, one will improve decision-making. In using the price system of a market, prediction markets are a structured approach for aggregating dispersed information from people.

2.1 The Price System as Information Aggregator

Consciously or subconsciously, there is a well-established mechanism for information aggregation in our society: the price system of a market. People (or speculators) keep buying and selling commodities in order to make profits. If a speculator knows that the price for a certain commodity will go up, he would buy today and sell tomorrow at a higher price. Vice versa, if he knows that the price will go down, he would sell today and buy tomorrow at a lower price. In both situations, he would have made a profit. To figure out where the price will go, speculators look for information or patterns that predict price movements. Then, they reveal this information to the market through the process of trading. In doing so,

speculators make the pattern go away as the information will be incorporated in the market price. For example, assume one finds that the price for a commodity generally increases on Tuesdays. Then, one would buy on Mondays and sell on Wednesdays. As a result, the price won't go up on Tuesdays as much any more. The more speculators compete in looking for information or patterns that predict price movements, the harder it becomes to find information that is not yet embodied in the market price. That way, the price system of the market is a powerful mechanism for aggregating information from people.

The idea that speculative markets perform well in aggregating dispersed information is long known and has been studied as the 'efficient market hypothesis'. Prediction market researchers often refer to this concept, which claims that such aggregation is 'perfect' in the sense that the prices 'fully reflect all available information at any time'. Yet, as outlined in Section 1.1, the prediction markets known to date are far from being efficient. Thus, this claim seems indefensible and prediction market researchers should refrain from citing it. That said, information aggregation in markets works well, even if efficiency cannot be proven with mathematical equations or equilibrium analyses. Rather, prediction market researchers should stay with Hayek (1945). In his article *The Use of Knowledge in Society*, published long before researchers studied the efficient market hypothesis, Hayek promoted the ability of the price system to aggregate information but never claimed that these prices meet the efficiency criteria of mathematical economics:

> *"I fear that our theoretical habits of approaching the problem with the assumption of more or less perfect knowledge on the part of almost everyone has made us somewhat blind to the true function of the price mechanism and led us to apply rather misleading standards in judging its efficiency."*
>
> (Hayek 1945, p.527)

2.2 The Concept of Prediction Markets

Prediction markets utilize the price system of the market not for speculation but for the primary purpose of aggregating dispersed information from people. Thereby, in incorporating human judgment and translating it into a numerical estimate, prediction markets can be classified as a judgmental approach (see the *Methodology Tree of Forecasting* at ForPrin.com). The price system of the market aggregates qualitative information or "knowledge of the kind which by its nature cannot enter into statistics" (Hayek 1945, p.524).

The idea of prediction markets is to set up a contract whose payoff depends on the outcome of an uncertain future event. This contract, which can be interpreted as a bet on the outcome of the underlying future event, can then be traded by participants. As soon as the outcome is known, participants are paid off in exchange for the contracts they hold.

Figure 1: Operational Principle of Prediction Markets

(adopted from Graefe et al. (2009c))

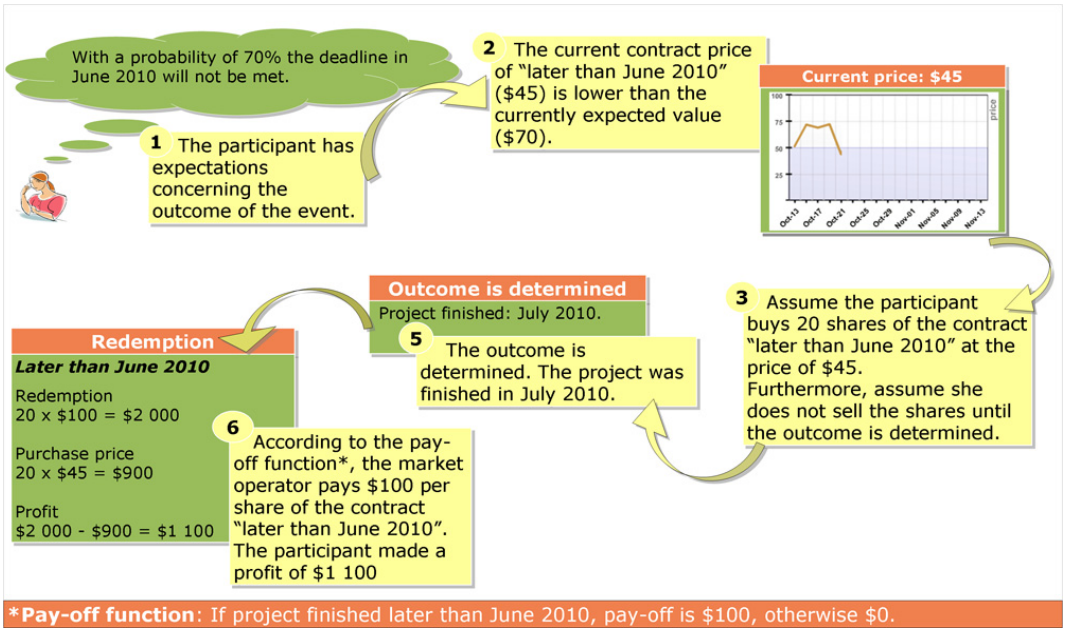


Figure 1 shows an example of how prediction markets work. This market aims at predicting whether the development of a new product will be finished by June 2010. The contract 'later than June 2010' pays off $100 if the product will eventually be finished later than June 2010; otherwise the pay-off is $0. This way the contract price can be interpreted as the probability of not meeting the project deadline in June 2010. If a participant believes with a probability of about 70% that the project deadline will not be met, she should be willing to buy (sell) these contracts for any price less (more) than $70. Thus she will become active if the current forecast differs from her individual assessment. Assume that the participant bought 20 shares of this contract (at a price of $45) and assume that the project was actually finished in July 2010. In this case, the participant would have made a profit of $1,100.

Table 1: Active contracts and predictions at intrade.com

(as of February 2009)

Contract	**Predicted probability of occurrence (order spread)**
Any country currently using the Euro to announce their intention to drop it on/before 31 Dec 2010	27-29%
EU agrees before end of 2009 to reduce CO2 emissions by 10% or more by year 2025	72-80%
Osama Bin Laden to be captured/neutralised by 30 Sep 2009	8-9%
Iran to conduct a nuclear weapons test on/before 31 Dec 2010	12-30%
Iran to be removed from US State Dept list of State Sponsors of Terrorism on/before 31 Dec 2009	13-15%
Steve Jobs to depart as CEO of Apple on/before 31 Dec 2009	55-57%
California Supreme Court to overturn Proposition 8 (gay marriage) by 31 Dec 2009	31-35%
A magnitude 9.0 or greater earthquake to occur anywhere on/before 31 Dec 2009	7-10%
Venue in North America to host the 2016 Summer Olympics	46-49%

The topics on which one can set up prediction markets are countless. To get an impression, Table 1 shows a small portion of contracts from the real-money prediction market intrade.com, along with the markets' probability estimates for the occurrence of the events. For example, at the end of February 2009, market participants agreed that, with a probability of about 46% to 49%, the 2016 Summer Olympics will be held in

North America. Now, anyone who thinks that these estimates are unreasonable can correct them by trading. If one thinks, the estimates are too low, one would buy stocks – and the price would go up. Vice versa, one would sell stocks – and the price would go down. As a compensation for the efforts to reveal information to the market, one expects to make profits. The claim is that – by not getting active in trading and aiming at correcting the market estimates – one agrees that these estimates are reasonable. In other words, there is an incentive to search for – and reveal – information whenever one thinks the current estimates are inaccurate.

Prediction markets are not limited to probability estimates but can also be used to predict absolute or relative numbers. Depending on the type of information one is interested in, different contract types have to be used.

If, as above, the goal is to obtain probability estimates for an event to occur, one usually uses 'winner-take-all' contracts. Such contracts have become well-known for predicting election winners. They are designed in a way that they pay off $100 if a candidate wins the election and $0 otherwise. Then, a contract price of $60 for 'candidate A' can be interpreted as a 60% chance for this candidate to win the election.

If the goal is to predict absolute or relative numbers, one uses 'index' contracts, which link the payoff directly to a number. They have also been used in election forecasting to predict vote shares. Such index (in election forecasting also known as 'vote-share') contracts are designed in a way that they pay off $1 times the actual vote share of a candidate. For example, if the received vote share is 44.4%, the market would pay off $44.4.

Both contract types are utilized in the IEM's U.S. presidential election markets: a winner-take-all market to predict the election winner and a vote-share market to predict the popular two-party vote shares of the Republican and Democratic

candidates. For a further discussion of the different contract types, see Wolfers and Zitzewitz (2004) or Luckner (2008).

2.3 Evidence on Accuracy

The performance of prediction markets has been analyzed for various fields of application. For an overview and categorization of research in the field see Tziralis and Tatsiopoulos (2007) or MIT's *Handbook of Collective Intelligence*.[12] This section summarizes some of the main research findings regarding the accuracy of prediction markets.

2.3.1 Election forecasting

Most of the work on using prediction markets for election forecasting has been done for predicting the outcome of U.S. presidential elections.

- In comparing IEM forecasts to 964 polls for the five U.S. presidential elections from 1988 to 2004, Berg et al. (2008a) found that the respective market forecasts were closer to the actual election results 74% of the time. This performance was replicated for the 2008 election (Berg et al. 2008b).

- Jones (2008) analyzed the forecasts of IEM's vote-share market for the 2004 election and compared them to traditional polls, a Delphi expert survey, regression models and a combination of all four approaches: the Pollyvote. He concluded that, in comparison with these methods of forecasting the popular vote, the IEM was the superior performer.

[12] The *Handbook of Collective Intelligence* is licensed under a *Creative Commons License* and available at http://scripts.mit.edu/~cci/HCI/index.php?title=Main_Page#Prediction_markets

- This performance was similar in 2008. In comparing the last forecasts before Election Day, Graefe et al. (2009a) found that the IEM was more accurate than polls, an Expert Delphi, the average of 16 quantitative models as well as the PollyVote. For more information see www.pollyvote.com.

However, the markets' advantage over polls disappeared when comparing the market forecasts to 'damped polls'. Damped polls involve historical information and re-calculate poll results based on the outcome of past elections. In analyzing data for the five U.S. presidential elections from 1988 to 2004, Erikson and Wlezien (2008) found these damped polls to be more accurate than both the IEM's vote-share and winner-take-all markets. For a further discussion of the relative performance of prediction markets and polls see Stix (2008).

2.3.2 Sports forecasting

Several studies have analyzed the performance of prediction markets for predicting the outcome of sports events.

- For predicting the results of 208 NFL games, Servan-Schreiber et al. (2004) compared the forecasts of two markets (one play-money and one real-money market) to those of 1,947 self-selected individuals. At the end of the season, the markets ranked 6th and 8th compared to the individuals. By comparison, the human average – which would be the outcome of a classical survey – ranked 39th.

- Spann and Skiera (2009) examined the accuracy of prediction markets for predicting the outcomes of German premier soccer league games and compared them to the performance of single experts (tipsters) and betting odds. In analyzing data from 678 games, they found that prediction markets performed equally well to betting odds and were clearly more accurate than tipsters.

- Luckner et al. (2008) compared the forecast accuracy of a prediction market for the FIFA World Cup 2006 to predictions derived from the FIFA world ranking and to a naïve model (random predictor). They found that the FIFA World Cup prediction markets were more accurate than both benchmarks. Luckner (2008) also compared these market forecasts to betting odds. In showing that the prediction market forecasts were equally accurate to betting odds, the findings were similar to Spann and Skiera (2009).

2.3.3 Business forecasting

Some small-sample studies have been conducted to analyze prediction markets for business forecasting.

- For forecasting sales figures, Chen and Plott (2002) reported on an internal market at Hewlett-Packard that beat the official forecasts of the company in 6 out of 8 events.

- Spann and Skiera (2003) compared forecast accuracy of an internal market at a large German mobile phone operator. They found that the market forecasts were more accurate than four extrapolation models (arithmetic mean, geometric mean, linear trend and exponential trend) for forecasting the use of five mobile phone services. However, due to small samples, these differences were not significant.

- In an early application, Ortner (1998) reported on an internal market at Siemens, where market participants correctly predicted the delay of a software project three months before the scheduled deadline. An article published in *IEEE Spectrum* reported similar findings for an internal market at Microsoft (Cherry 2007).

2.3.4 Other applications

In addition, several other studies have been published that analyzed the performance for different field of application and different types of problems.

- For predicting Oscar Award winners, Pennock et al. (2001b) compared predictions from the 'Hollywood Stock Exchange' (HSX)[13] to expert forecasts of five movie columnists. On the day the experts revealed their forecasts, one of them was better than the market predictions. From the day after, the market outperformed all experts as well as the average of the experts' estimates.

- Pennock et al. (2001a, p.987) analyzed results from the Foresight Exchange (FX), a play-money prediction market that aims at predicting events in the far future. For 161 contracts that referred to 'yes' or 'no' questions, the researchers recorded the FX forecasts thirty days before the respective outcome was known. They found that the FX forecasts strongly correlated with outcome frequencies but did not compare the results to a benchmark forecast.

- Soukhoroukova (2007) used prediction markets to evaluate new product ideas as well as new product concepts. For evaluating new product concepts, she found that the markets provided results that were consistent with participants' self-explicated expectations and traditional methods in product planning, like conjoint analyses. Yet, in the case of evaluating new product ideas, the market results did not reflect the assessments of an expert panel. In a similar study, Dahan et al. (2007) also used prediction markets to assess new product concepts and found high consistency of market results with the results from an independent survey study.

[13] http://www.hollywoodstockexchange.com

2.3.5 Summary

In sum, the available studies demonstrated high forecasting accuracy and showed prediction markets to perform at least as well as alternative forecasting methods. These include (individual and aggregated) expert opinions, polls, betting odds, quantitative methods as well as naïve benchmark models. However, the number of studies available to date is still limited and they are often small scale. Thus far, no study has been published that provided a meta-analysis of prediction markets' accuracy. There is a remaining need for further empirical studies.

2.4 Promising Features

Prediction markets possess several characteristics that appear to be beneficial to improve on existing forecasting methods. They can enable continuous and ad hoc aggregation of new information and can motivate participation. In addition, they are a cost-efficient, scalable and feed on multiple sources of information. For a thorough discussion of the potentials of prediction markets see Graefe et al. (2009c).

2.4.1 Enhancing quantitative forecasting methods

Quantitative forecasting methods often rely solely on historical data, assume for stable conditions, and are therefore not able to deal with unexpected changes. By comparison, prediction markets are a judgmental forecasting method that motivates participants to feed on all available sources of information. This can involve historical data, forecasts from other approaches, news as well as individual expectations. Furthermore, since participation is only beneficial if one does *not* agree with – and is able to improve – the current market forecast, prediction markets trick participants into constantly challenging the group opinion and to actively search for superior information. Thus, prediction markets can motivate participants to actively think about the future.

2.4.2 Continuous and real-time information aggregation

Often, forecasting methods are conducted as one-off activities at a single point in time and, thus, have a rather episodic character. This is because the results of traditional approaches (like quantitative methods or Delphi studies) usually have to be manually analyzed, evaluated, and summarized by a facilitator.

In contrast, the price mechanism of the market automatically aggregates all available information that is dispersed among participants. Thus, prediction markets can reflect the aggregated group opinion *at any time*. This has two positive effects. First, automatic information aggregation reduces the workload of the operator. Second, the market enables real-time and continuous information aggregation since price changes immediately react to the availability of new information. Thus, the implications of sudden events are quickly incorporated in the forecast. As a result, the availability of results is not tied to certain points in time but the market functions like a dynamic system that reacts to new information revealed by participants.

2.4.3 Motivating information revelation

Depending on their forecasting or trading performance, participants can win or lose money. Therefore, by the anticipation of profit, markets motivate participation and revelation of *information* – instead of preferences. In providing performance-based incentives, prediction markets can also overcome problems of group tendencies or groupthink (Janis 1972). Group tendencies often occur in traditional consensus development methods like Delphi and entice participants to follow the herd. For example, in reviewing the literature, Woudenberg (1991) found that consensus in Delphi is mainly achieved by group pressure to conformity or attrition of participants.

Cherry (2007) reported on an internal market implemented at Microsoft that illustrated how prediction markets can overcome group pressures. The goal of this market was to

predict the launch date of a certain software product. After trading has started, the market instantly predicted that the product will not be finished in due time. Thus, the markets revealed information that no member of the developer team has directly communicated to the project manager before – evidently, a strong example of group pressures. The management trusted the market forecast and cut some features that were considered to slow down the development process. In turn, this decision was again immediately reflected by the market, indicating higher probabilities that the product might still be finished in time. In the end, when the customers demanded the features back, the market again predicted that the product would be finished late – and was finally right.

2.4.4 Motivating participation

Cuhls (2003) argued that it can be challenging to motivate participation in traditional approaches like Delphi for ongoing forecasting activities, especially over long periods of time.

In providing performance-based incentives, prediction markets can be used to motivate participation. This can be done in various ways. For example, the market operator can award prizes or money to the best participants or the participants can be remunerated depending on their portfolio value. For a discussion of when to use different pay-off mechanisms in play-money markets see Luckner and Weinhardt (2007).

However, it is not essential to provide monetary incentives to motivate participation. For example, in internal markets at Google, participants did not seem to care about cash prizes but wanted to know about reputational prizes like shirts that would identify them as winners (Graefe 2008e). Furthermore, by simply announcing a user ranking, the play-money markets in Christiansen's (2007) field experiment performed well without providing monetary or tangible incentives at all.

Often, the social incentive of reputation and the possibility to match someone's personal expertise with others is sufficient to motivate participation. This incentive of social competition may be even stronger within an organization where people know each other. In addition, there are concerns that providing monetary incentives may actually be counterproductive in internal corporate markets if the incentives are large enough to entice employees to work against the goals of the organization. For a further discussion of how to design incentive mechanisms to motivate trader participation see Graefe (2008e).

2.4.5 Scalability and cost-efficiency

Due to time or money restrictions, traditional judgmental forecasting approaches like Delphi or nominal groups are limited in their number of participants. As described above, information aggregation in prediction markets is carried out automatically via the price mechanism. No manual intervention by a facilitator is required, which significantly reduces the workload at runtime. This makes prediction markets arbitrarily scalable as the workload does not increase with the number of participants or the time horizon of the study. As argued by Spann et al. (2007), the hardware costs for running the market are negligible once the market platform has been designed.

2.4.6 Participatory regulation

As summarized in Section 2.3, the track record of prediction markets is good. For various types of problems, they have been shown to perform at least as well as alternative forecasting methods. However, the nature of most of the problems for which prediction markets have been used is simple: they only require aggregating information or 'facts' and its outcome does not relate to complex decisions being made. Examples include forecasting elections, sports events, or sales figures.

In motivating participation and information revelation and aggregating dispersed information from a – virtually unlimited – number of people, prediction markets have the potential to provide social utility. They can contribute to participatory regulation and aid

the democratic process by incorporating the views of a broad public on problems that ask for policy interventions. Such problems are complex as they not only require aggregating information or facts. They also involve people's values, attitudes, emotions, expectations, fears, or commitments. For such problems, achieving consensus becomes highly difficult.

Graefe and Orwat (2007) discussed the use of prediction markets for participatory regulation and compared them to traditional mechanisms of public engagement. They found that, in aggregating information from preferably large, arbitrarily composed samples in a structured manner, prediction markets enable a new and unique type of public engagement mechanisms. Unfortunately, also due to the cancellation of the PAM (see Footnote 4), thus far there is no empirical research that analyzed the use of prediction markets for suchlike problems. In an effort to address this need, Graefe and Armstrong (2009b) have recently started a project to examine the use of prediction markets to solve complicated public policy issues like climate change.

Chapter 3

Research Methodology

This chapter describes the research methodology. First, the three benchmark approaches, to which prediction markets were compared in this work, are described. Then, a categorization of task types is provided. Finally, the experimental settings of the TechForX field experiment as well as the laboratory experiment on group technique comparison are presented.

3.1 Group Techniques for Information Aggregation

The two empirical studies in this work compared prediction markets to three other group techniques for information aggregation, all of which differ in the amount and structure of interaction permitted between group members. The three group techniques were face-to-face meetings, nominal groups, and the Delphi method.

3.1.1 Face-to-face meetings

Face-to-face meetings (FTF) are the traditional and most common approach for group decision-making in organizations, allowing any form of direct interaction between group members. In this context, the goal of this unstructured discussion is to achieve a final group

estimate for a problem. Traditional meetings have been shown to be subject to many biases and drawbacks:

- It requires time and effort for a group to maintain itself (Dalkey & Helmer 1963).
- Groups tend to aim at reaching 'speedy decisions' and to not consider all problem dimensions (Maier & Hoffman 1960).
- Groups tend to pursue a limited train of thought, which leads to a 'central tendency effect' or 'groupthink' (Janis 1972).
- Less confident group members, or people from lower hierarchy levels, may silent themselves because of group pressures for conformity or implied threats of sanctions (Dalkey & Helmer 1963).
- Dominant personalities tend to exert excessive influence on the group (Dalkey & Helmer 1963).
- A 'self-weighting' effect occurs: group members try to participate and to exert influence to a level that they feel equally competent with others (Kelley & Thibaut 1954).

For thorough review of these issues see Van de Ven and Delbecq (1971). In sum, Armstrong (2006) found little evidence to support the use of meetings for forecasting or decision-making. In addition, meetings are expensive to schedule and to run.

In some situations, however, one must be concerned not only with the quality of a decision but also with its acceptability. Personal interaction in groups can either lead to coherence, and thus high perceived satisfaction, or disagreement, resulting in frustrated group members. Yet, in general, people enjoy human interaction and the sense of working together and meetings have been shown to often achieve high levels of satisfaction (e.g. Boje & Murnighan 1982, Van de Ven & Delbecq 1974).

3.1.2 Nominal group technique

The nominal group technique (NGT), developed by Van de Ven and Delbecq (1971, 1974), tries to account for some of the drawbacks of traditional meetings by adding a structured format to direct interaction.

This process is conducted in three steps: First, group members work independently and generate individual estimates on a problem. Then, the group enters unstructured discussion to deliberate on the problem. Finally, group members work again independently and provide their final individual estimates. The group result is the aggregated outcome of these final individual estimates. The literature refers to this process also as *estimate-talk-estimate* (Gustafson et al. 1973).

The idea of NGT is that direct interaction during the assessment or evaluation phase can have a positive impact on problem solving. In particular, it can help group members to clarify and justify their point of views which may help the group to make more informed decisions. Nonetheless, in the phases of generating answers and making the final decisions, NGT prevents direct interaction between group members to reduce the drawbacks known with traditional meetings. Van de Ven and Delbecq (1971) summarized some of the advantages of nominal groups:

- In incorporating direct interaction and presence of others, NGT provides evidence of activity and retains the social facilitation of the group process.
- It eliminates evaluation or elaborating comments when generating the problem dimensions.
- It provides participants with time for reflection and forces them to record their thoughts.
- It limits the influence of dominant personalities on the group outcome by involving the judgments of all group members.

3.1.3 The Delphi method

Unlike FTF or NGT, which require physical proximity of group members, participants in Delphi are physically dispersed and do not meet in person. In general, Delphi functions somewhat similar to NGT. The main difference is that written interaction is utilized during the whole process to prevent any form of direct interaction between group members. The idea is to completely eliminate social biases due to direct interaction by keeping participants anonymous.

3.1.3.1 The concept of the Delphi method

The Delphi method, in the following simply referred to as Delphi, was developed by the RAND Corporation in the late 1950s. As described by Dalkey and Helmer (1963), the goal of its original implementation was to apply expert opinion to the selection of an optimal U.S. industrial target – from the viewpoint of a Soviet strategic planner.

Delphi is a multiple-round survey in which participants anonymously reveal their individual estimates as well as comments on a problem. After each round, the individual estimates and comments are summarized and reported as feedback to participants. Taking into account this information, participants provide their new estimate in the following round. The group result is the aggregated outcome of the individual estimates of the final round. The strengths of the method are seen in its structured communication process that enables discussion and helps a group achieving consensus but limits the drawbacks associated with direct interaction. For reviews of Delphi see Woudenberg (1991) and Rowe and Wright (1999, 2001).

3.1.3.2 The use of Delphi in practice and science

Delphi has been widely used by businesses, governmental agencies, and organizations to derive predictions of future developments. In most cases, Delphi was used for business forecasting. Examples include the Argentine power sector, broadband connections, dry bulk shipping, leisure pursuits in Singapore, rubber processing, Irish specialty foods or oil prices.

Forecasts of technology were also popular. These included forecasts about intelligent vehicle highway systems, industrial robots, intelligent internet, and technology in education. Other applications were concerned with broader social issues such as the 'urban future' of Nanaimo in British Columbia or the future of law enforcement.

Using the keywords 'Delphi AND (predict OR forecast)', searches of the *Social Sciences Citation Index* and the *Science Citation Index Expanded* were conducted to assess what has been happening to researcher interest in Delphi over the years. Altogether, 65 relevant items were identified; 1 from the 1960s, 8 from the 1970s, 3 from the 1980s, 21 from the 1990s, and 32 so far this decade. Even though the method is around since the 1950s, it appears as if researcher interest in Delphi is constantly increasing. For more information on the use of Delphi in research and practice see Green et al. (2007).

3.1.3.3 Characteristics of Delphi and prediction markets compared

In meeting the four key features of Delphi (Rowe & Wright 1999), prediction markets work to some extent similar:

1. Both are structured approaches that use a set of predefined hypotheses.
2. These hypotheses are judged anonymously.
3. Both methods incorporate a feedback mechanism. However, the nature of feedback differs. While both provide numerical estimates, Delphi participants can also reveal reasons and comments for their estimates.[14]
4. The goal of both mechanisms is to achieve an aggregated group response.

[14] Note that it is technically possible to provide additional means of communication in prediction market (like chats, blogs, etc.). However, to date, no research is known that analyzed the effect of such means of communication on the performance of prediction markets.

The close relationship of Delphi and prediction markets is also illustrated by the *Methodology Tree of Forecasting* at ForPrin.com.

Despite their similarities, to date there is no empirical evidence on the relative accuracy of both methods. However, prediction markets have the potential to avoid some potential drawbacks of Delphi:

- While Delphi can help group members reach an understanding, Woudenberg (1991) found that such an understanding can arise as a result of group pressures to conformity and, thus, may conceal disagreement. By comparison, participants in markets can only benefit if they challenge the group opinion. As a result, they should be motivated to second-guess their own judgments and to reveal their true beliefs.

- Delphi is typically conducted as a one-off activity, providing results at a certain point in time. Although new forms like Real-time Delphi (Gordon & Pease 2006) enable live feedback, the question remains how panelists can be motivated to participate over longer periods. In prediction markets, incentives can motivate people to keep participating. Furthermore, markets can be run continuously so that new information is instantly and automatically incorporated into the forecast.

- It can sometimes be difficult to recruit suitable experts to participate on a Delphi panel. In the case of prediction markets, participants themselves choose to take part if they think that their private information is not yet incorporated in the forecast.

By comparison, Delphi seems to have these advantages over prediction markets:

- Delphi can be used for a much broader range of problems, since there is no need to judge the outcome of a situation in order to determine pay-offs for participants.

- As will be outlined in Sections 3.3.2 and 3.4.2, it can be challenging – if not impossible – to formulate some problems as contracts in prediction markets. It is easier to address complex issues and to obtain predictions by asking direct questions in using questionnaires.

- Many people lack the understanding of how markets work or how to translate their expectations into market prices. It is easier for people to reveal their opinions in Delphi.

- It is easier to maintain confidentiality with Delphi. As has become evident with the cancellation of the PAM (see Footnote 4), it may be morally objectionable to benefit from trading on the outcome of critical issues. As described in Section 1.2.3, concerns may also arise over the use of markets within businesses or where the forecast may de-motivate employees.

- The opportunity to provide comments or reasons for judgments allows Delphi participants to introduce new ideas into the discussion. Also, the transparent exchange of knowledge allows experts to learn while participating in the Delphi process.

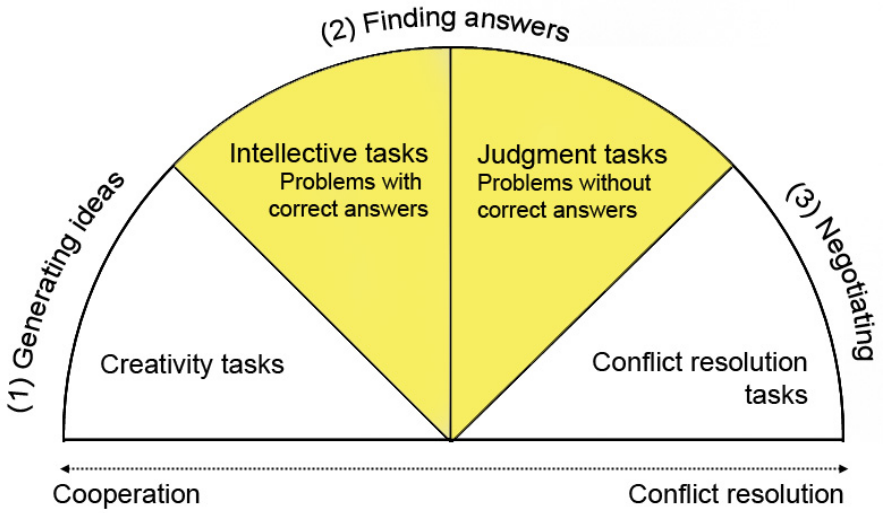
In sum, by providing incentives for participation and automatically aggregating information, prediction markets can be expected to have some advantages over Delphi. In particular, markets should be superior for short-term problems that are easy to address, when it is possible and desirable to involve a large number of participants, and when the primary goal is obtaining a group response. In contrast, the uses of Delphi are not limited to providing reliable and accurate forecasts. Delphi is also valuable because it can be used in situations of high uncertainty by enabling communication, encouraging discussion, and generating new ideas. For a further discussion of the relative advantages and disadvantages of Delphi and prediction markets see Green et al. (2007).

3.2 Group Judgment Tasks

Findings from group studies are generally task-specific and cannot be transferred to tasks of a different type. As can be seen in Figure 2, McGrath (1984) distinguished three categories of task types: (1) generating ideas (creativity tasks), (2) finding answers (intellective or judgment tasks), and (3) negotiating (conflict resolution tasks). These task types differ in terms of interaction as the interdependencies between group members successively increase.

Figure 2: Classification of Cognitive Task Types

--adopted from McGrath (1984)--



The hemisphere in Figure 2 can be separated into two halves. The tasks in the left half (creativity and intellective tasks) tend to be cooperative. For example, in using simple brainstorming techniques, *creativity* tasks ask people to reveal ideas, without having the quality of these ideas evaluated by others. Thus, in the process of generating as many ideas as possible, consensus among – or feedback from – group members is not required. Rather, evaluative and emotional components are discouraged. On the other end of the spectrum, *negotiation* tasks are competitive and most challenging for achieving consensus. The reason is that such tasks do not only require aggregating information or 'facts'. They also involve

people's values, attitudes, emotions, expectations, or commitments. For such tasks, achieving consensus becomes highly difficult.

Within both extremes lie tasks that require participants to find answers for certain problems. As described by Laughlin (1999), such tasks can be ordered on a continuum that is anchored by *intellective* and *judgment* tasks.

Intellective tasks ask group members to solve problems that have demonstrably correct answers. Thereby, demonstrability of the correctness of the answers varies. For some problems, the correctness of an answer can be demonstrated at the time the judgment is being made. Once the answer is determined, it is obvious to every group member. The literature refers to such problems also as *Eureka* problems. Examples are logic problems (e.g. puzzles) or algebra problems. Thus, if one knows the answer for such problems, one should have little difficulty to demonstrate its correctness and, thus, to convince others to adopt the answer. Accordingly, it is easy for a group to achieve consensus on such an answer. The criterion of successful group performance is the achievement of the correct answer.

Judgment tasks ask for evaluative, behavioral, or aesthetic judgments and do not have demonstrably correct solutions. Examples are jury decisions or matters of taste, like the evaluation of new products concepts. Thus, the basic goal of a group is to achieve consensus.

The two experiments reported on in this work analyzed two different types of tasks. In the TechForX field experiment, described in Section 3.3, participants had to solve a classical judgment task, consisting of questions without demonstrably correct solutions. In the laboratory experiment on the relative performance of prediction markets, described in Section 3.4, participants had to solve a variation of an intellective task, which will be referred to as a quantitative judgment task. These tasks asked participants to reveal percentage estimates on factual questions. While these questions had correct solutions, participants could not verify the correctness of answers at the time of the judgment. Thus, all group

members, as well as the group as a whole, could be expected to have some uncertainty about their answers. Therefore, the ability of the group to efficiently aggregate the relevant information of its members was crucial to its performance.

3.3 TechForX – A Field Experiment on Long-term Forecasting

Part of this work reports on findings from the TechForX (Technology Foresight Exchange) field experiment that was conducted in line with the European Foresight Project EPIS[15]. The goal of TechForX was to compare prediction markets to the Delphi method for long-term forecasting problems.

3.3.1 Study design

In its first period from 2006 to 2007, the EPIS project focused on assessing future technological trends in the creative content industry[16]. In order to obtain views from a

[15] EPIS (European Perspectives on Information Society) was a multi-annual project, run by the Institute for Prospective Technological Studies of the European Commission's Joint Research Centre, on behalf of the European Commission's Information Society and Media Directorate-General. The project aimed at observing trends in technology and business models of Information and Communication Technologies (ICT). In running foresight exercises on selected areas in ICT, it should investigate specific implications for policy (in particular research and development policies in ICT). For further information on EPIS visit the project website at http://epis.jrc.ec.europa.eu.

[16] In the context of the EPIS project, the Creative Content Sector was defined as the collection of activities involving the 'creation and distribution of goods with an intrinsic cultural, aesthetic or entertainment value which appears linked to their novelty and/or uniqueness'. This definition made it possible to adopt a tolerant characterisation of the subject matter, avoiding traditional differentiations between "high" and "low" cultural activities. It also provided a clear separation between creative content and communication media industries such as magazines, television or news broadcasting. Based on this definition, the following sub-sectors of the Creative Content sector were included in the EPIS analyses:

- Audiovisual production (film and television broadcasting, excluding advertising)
- Music Recording and Publishing

potentially broad audience, the core project team conducted a Delphi study. In parallel, the opportunity arose to launch the TechForX prediction market as an experimental add-on. Implementing TechForX in this environment allowed examining the applicability of prediction markets for long-term forecasting in a real-world setting with relevance to political decision-makers. Table 2 gives an overview of the study design.

Table 2: TechForX study design overview

	EPIS Delphi	TechForX prediction markets	
		Experts	Students
Runtime in weeks	4	2	2
No. of theses	36	5	5
No. of participants invited	1,111	25	25
No. of participants responded	288	18	20
Monetary incentive	No	Yes	Yes

3.3.1.1 EPIS Delphi

The EPIS Delphi was conducted as a so-called *Real-time (RT) Delphi*. The RT Delphi is a completely web-based version of the traditional Delphi in which the round-based procedure is abandoned. Instead, participants can change their opinion at any time and as often as they like, whereas the aggregated group opinion is recalculated whenever a participant changes his opinion. This is expected to lessen some weaknesses of traditional Delphi, which is considered to be time-consuming and resource-intensive. Furthermore, by processing information in real-time, it makes the approach more similar to prediction markets. For a discussion of RT Delphi see Gordon and Pease (2006). For a detailed report on implementation and results of the RT Delphi within EPIS06 refer to Friedewald et al. (2007).

- Printing and Publishing (focusing on Books)
- Video game development and publishing
- Cultural spaces (museums and libraries)

For a further discussion and definition of the understanding of the Creative Content Sector see Mateos-Garcia et al. (2007).

3.3.1.2 TechForX prediction market

TechForX was based on the STOCCER market platform that has been used to predict the outcome of the FIFA World Cup 2006 and German premier league soccer games. For a thorough description of the STOCCER market platform see Luckner (2008). TechForX has been implemented as a play-money prediction market using a continuous-double auction (CDA) in combination with limit orders. In a CDA, traders place buy (or sell) orders up to a specific volume and price. Trades occur if prices match or if the price of a buy order exceeds the price of a sell order.

In a play-money market, participants do not invest real money but are endowed with an initial amount of virtual money and stocks. Every participant received 10,000 virtual currency units per market as well as 100 shares of every contract in the particular market. In order to ensure circulation of shares, participants could buy and sell so called 'unit portfolios'. A unit portfolio consisted of one share of each contract available in the market and could be purchased from or sold to the market system at any time. The portfolio price was determined by the aggregate pay-off for one share of every contract in the market and always equaled 100 virtual currency units in each of the TechForX markets. While portfolio trading was completely risk free for participants, the net purchases of unit portfolios were used to endogenously determine the supply of contracts.

3.3.2 Task type

Participants in the Delphi study and the prediction markets had to judge the realization time of several theses that focused on the development of the creative content sector. These theses aimed at predicting developments in the far future. Accordingly, the correctness of answers could not be verified at the time of judgment, neither by participants nor by the project team. Participants had to solve a classical judgment task (cf. Section 3.2).

3.3.2.1 Development of theses

The theses were developed by the EPIS project team. For that purpose, the project team invited experts from the field and conducted workshops to discuss how the creative content sector might develop in the future. For a detailed description of the process for generating the theses see Friedewald et al. (2007). Finally, the project team came up with a list of 36 theses.

3.3.2.2 Generation of the Delphi questionnaire

Delphi participants were asked to assess the time of realization for each of the 36 theses. Thereby, they had to select one of seven predefined time horizons:

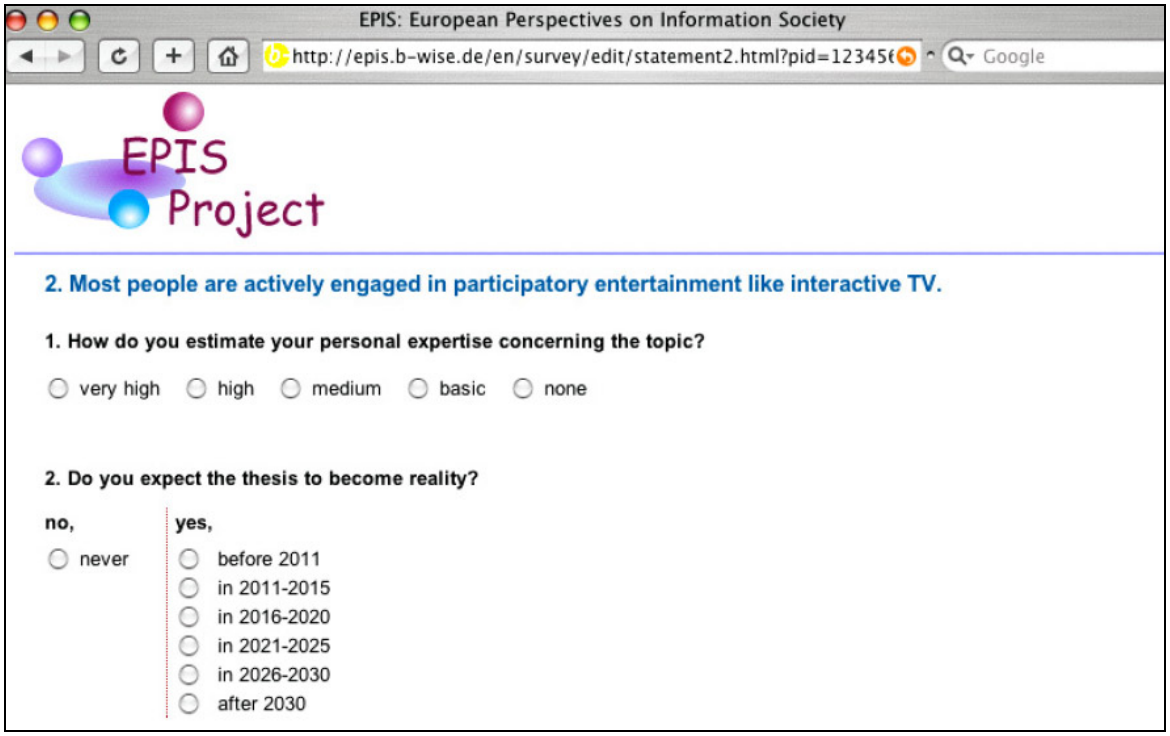
(a) 'Up to 2010'
(b) '2011-2015'
(c) '2016-2020'
(d) '2021-2025'
(e) '2026-2030'
(f) 'Later'
(g) 'Never'

Thus, participants gave binary answers by assigning 100% to the time horizon they think is most likely and implicitly – and automatically – assigning 0% to all others. Then, the median was applied to calculate the aggregated group opinion as the probability distribution over all answers. An example of the layout of the online Delphi questionnaire is shown in Figure 3.

3.3.2.3 Mapping the Delphi to prediction markets

In TechForX, Delphi theses were implemented as markets and time horizons were mapped as contracts. To obtain the group opinion as a probability distribution from a prediction market, it would have been necessary to map each of the seven time horizons as a single contract. This would have led to seven contracts per market. For 36 theses, this would have led to 252 contracts. Yet, this would have yield usability problems since information revelation in TechForX was more challenging and time-consuming compared to the Delphi.

Figure 3: Layout of the EPIS Delphi Online Questionnaire



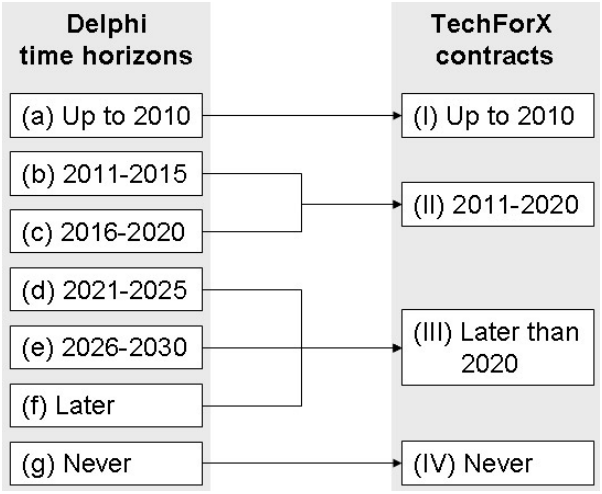
The reason is that CDA markets do not ask participants to reveal binary answers but their whole probability distribution. For example, assume the price for the time horizon 'Up to 2010' reflects a probability of 20% for a particular thesis to come true. If a participant thinks this time horizon is undervalued (assume he thinks it is 25% likely that this thesis will come true 'Up to 2010'), he should buy shares. At the same time, provided the whole market has been in equilibrium, at least one other time horizon must have been overvalued which means that he should have sold shares of those time horizons. Thus, prediction markets motivate participants to cogitate about their individual probability distribution over all time horizons. In other words, participants have to reveal information more precisely. However, it has to be considered that transferring one's opinion into market prices requires certain cognitive abilities as well as an understanding about the functionality of markets. As described in Section 3.1.3.3, information revelation in markets is far more complex and time-consuming as it is in Delphi.

Table 3: Theses and number of answers per thesis obtained in Delphi

Thesis no. (as referred to in this work)	Thesis (topic)	Thesis (verbalization)	No. of answers obtained in Delphi
1	E-Books	More than 20% of all books are purchased in electronic form and not as printed copies.	164
2	Interactive TV	Most people are actively engaged in participatory entertainment like interactive TV.	167
3	Presence in virtual worlds	More than 20% of all citizens are present in at least one virtual world, where many people are online and interact with each other (like Second Life).	106
4	3D-displays	3-dimensional displays are most common for playing online games.	99
5	DRM and P2P	When media is distributed over peer-to-peer (P2P) networks, normally Digital Rights Management is used.	109

For the purpose of this experiment, the markets' complexity was reduced by focusing on a limited number of 5 theses. These theses were selected by relying on independent ratings of two colleagues. As for seven time horizons this would still have led to 35 contracts, the number of contracts was reduced by combining certain time horizons. Figure 4 shows how the Delphi time horizons were mapped to contracts on the TechForX prediction market. Finally, participants were asked to assess four contracts per market – or 20 contracts altogether. Table 3 lists the five theses that had to be answered by both groups as well as the overall number of answers per thesis obtained in the Delphi.

Figure 4: Mapping EPIS Delphi to TechForX



3.3.3 Participants

3.3.3.1 EPIS Delphi

In the process of recruiting participants for the EPIS Delphi, addresses from public databases, journals, and conference proceedings were collected. In addition, the study was announced on relevant mailing lists and the foresight network was asked to recommend participants. The final database comprised 1,275 addresses, with most people affiliated with research organizations, enterprises, industry or innovation and ICT policy institutions.

Of the 1,275 e-mails that were sent, 164 (12.8%) were undeliverable. Of the remaining 1,111 experts, 426 visited the Delphi website and 288 of them started to fill out the questionnaire. Thereof, 124 returned to the questionnaire at least once in order to review the results or change their estimates.

Three-fourths of the respondents were from EU countries, while almost one quarter of the respondents were from EU non-member states. Most of the participants were from Germany, France, the United Kingdom and the United States (52 % altogether). The geographic distribution of the respondents was similar to the distribution of the original addresses. For further information on the recruitment of the Delphi participants see Friedewald et al. (2007).

3.3.3.2 TechForX

For TechForX, two groups of participants were recruited: one expert group and one control group. Each group consisted of 25 people. Participation was completely voluntary and each participant indicated his willingness to participate before the start of the experiment. It was ensured that none of the market participants concurrently participated in the EPIS Delphi.

The recruited experts were identified through web searches or were recommended by researchers from the foresight network. They were experienced either specifically in the

creative content sector or with foresight (or futures studies) in general. To ensure that the experts bring along a broad range of dispersed knowledge, the goal was to build up a diverse group in view of nationality, education, and occupational field. The 25 experts came from 11 different countries: 10 from Germany, three from Belgium, two from Australia, Netherlands, and the UK, as well as one from Greece, Peru, Russia, Switzerland, Taiwan, and the US. Most of them were senior researchers whereof seven held a doctoral degree. Another seven experts have been writing their PhD thesis at the time the experiment was conducted. Furthermore, two experts – one of them a science-fiction author – have been operating blogs that deal with the future of the content sector or foresight, respectively. The remaining experts were working either in the content industry or were associated with relevant policy institutions.

The control group – in the following referred to as the student group – consisted exclusively of freshmen studying information science at the Universität Karlsruhe (TH), Germany. Due to the subject of their studies, they were expected to bring along a certain degree of knowledge and interest for new media or technological developments related to the creative content sector.

3.3.4 Incentive mechanism

The results of the EPIS Delphi were used to determine the pay-off function of the markets. This means, at market closure, the contracts were liquidated at the probabilities derived from the Delphi study. For example, if the Delphi results predicted a probability of 25% for a particular thesis to come true 'Up to 2010', the respective contract was worth 25 at market closure. For the time horizons '2011-2020' and 'Later', which have been combined for the markets, the means of the respective Delphi results were used to determine the final contract values.

Then, for each trader, the overall deposit value for the 20 stocks in the 5 markets was calculated. Based on these deposit values, a final user ranking was calculated for both the

expert and the student market. Shown to be superior for remunerating participants in play-money prediction markets (Luckner & Weinhardt 2007), the rank-order tournament incentive scheme was used for defining the pay-off function. The three best-performing participants of each market were remunerated: 200€ for ranking 1st, 100€ for 2nd and 50€ for 3rd.

3.3.5 Materials and procedures

With the start of the experiment, a short instruction explaining the goal of the experiment, the functioning of the market, as well as the pay-off mechanism was sent to every participant via e-mail. In addition, this information was posted on the TechForX experiment website. The instruction is provided in Appendix M.1.1. Furthermore, participants could access a full tutorial of the TechForX platform (cf. Appendix M.1.2). During the runtime of the experiment, user rankings were provided for each of the five markets based on the last traded market price.

The EPIS06 Delphi was conducted between May 31st and July 2nd, 2007. The markets were open for trading on the internet for two weeks from June 22nd to July 6th. Although they were still running for four days after the Delphi was closed, no market participant had access to the Delphi results as they were not published before mid-August.

3.3.6 Participation statistics

Table 1 provides an overview of the trading activity of both groups of participants. Of the 25 individuals that have been invited for each group, 20 students and 18 experts actually participated in the experiment (i.e. each of those individuals has placed at least one order in any of the five markets).

Table 4: TechForX trading activity

	Students	Experts
Invited	25	25
Active	20	18
Number of Orders	1,672	403
Number of Trades	676	176

Although the number of active participants was nearly the same in each group, the number of orders / trades was about four times higher for students (i.e. students were about four times more active than experts). Several reasons might explain this difference:

- Compared to students, experts are usually busier and have further commitments. In fact, when inviting the experts, some of them already indicated that – although they were excited about participating – their available time may be limited. Considering time as an opportunity cost, one can assume that the prospect of winning 200€ was about four times more valuable for students than for experts.
- Many students may have known each other since the experiment was advertised in a lecture at the university. In contrast, as experts were invited independently and internationally dispersed, it is unlikely that a large number was acquainted with each other. Thus, the social incentive of climbing the user ranking might have been higher for the student group than for the experts.
- As has been shown by Tetlock (2006), experts are often very confident in their knowledge and insist on their judgments – even if they have been proved to be wrong. To some extent, this behavior may have also occurred in the expert group. Following Christiansen's (2007) classification of trader types, an analysis of trader behavior identified three initial-set traders and two one-stake traders among the experts. Among the students there was only one individual of each type. One-stake traders are individuals who make only one trade in each market in which they participate. Initial-set traders trade multiple stocks per market but never trade a stock twice. Participants belonging to both of these trader types reveal their opinion only once and do not reconsider based on the feedback provided by the group. This

trading behavior might already indicate higher confidence of experts in their individual judgments. The question about experts' and students' confidence will be further analyzed in Chapter 5.

In general, the number of orders and trades might seem low compared to, for example, election markets. For such events, new information becomes available continuously, for example due to certain actions of a candidate during the political campaign. On the contrary, for markets on assessing long-term developments, it is unlikely that new information (e.g. in response to real-world events) became available during the short runtime of the markets. In general, once the information of participants is incorporated in the market prediction, further trading is motivated only if participants disagree with the group opinion – or exploit arbitrage opportunities.

3.3.7 Market inefficiencies

In the finance literature, market analyses typically assume a no-arbitrage condition: if markets are perfectly efficient, the sum of the prices for each contract exactly equals the aggregated pay-offs at market closure. Thus, if the markets characterized here were perfectly efficient, the prices for the four contracts in each market should have always summed up to the unit portfolio price of 100 and, thus, met the no-arbitrage condition.

Arbitrage opportunities arise in a market whenever there are outstanding buy orders that sum up to a price of more than 100. In this case, any trader can realize a risk-free profit by accepting all four buy orders. Vice versa, traders can realize risk-free profits whenever the sum of the sell orders is less than 100.

In examining possible arbitrage violations at market closure, differences between students and experts were identified. While none of the five student markets showed arbitrage violations, experts did not exploit all possible arbitrage opportunities. In particular, in three

of the five expert markets, the sum of the buy orders at market closure was above 100. This led us to assume that experts neither cared about efficient markets nor about realizing small profits, which could already be expected after observing the low liquidity in the expert markets. Rather, being used to reveal their opinion in survey-based approaches like Delphi without getting compensation, experts seemed to look upon the market simply as what might be considered as its primary purpose: a tool for information aggregation.

Market inefficiency is a well-known and common problem in experimental settings with a small number of mostly inexperienced traders and prior research has shown that such markets do not meet the no-arbitrage condition. As in our markets, Rietz (2005) has conducted laboratory experiments in which arbitrage violations were consistent and led to absolute prices higher than rational price levels. Equally, Chen et al. (2004) reported on laboratory experiments with thin markets in which the no-arbitrage condition was only rarely met. Furthermore, arbitrage violations do not only appear in thin laboratory markets but have also been shown for the long-running Iowa Electronic Markets, which are populated by self-selected and, often, experienced traders (Forsythe et al. 1999, Oliven & Rietz 2004).

Although these inefficiencies do not prevent the market from aggregating information (Forsythe et al. 1999), they undermine the assumption of rational behaving participants. In consequence, the forecasts derived from market predictions may loose on reliability and credibility. Thus, it seems worthwhile to look for ways to minimize market inefficiencies. One way to do so is to open participation to traders providing additional liquidity to the market. For example, no inefficiencies were observed at market closure in the student markets. Arbitrage opportunities may disappear by inviting students to the expert markets whose primary goal may not have been information revelation but profit realization. Although one could worry about inviting uninformed participants, empirical studies showed that noise trading – or manipulation, respectively – does not harm accuracy (see Section 1.3). Another possibility is to replace the continuous-double-auction mechanism by an

automated market-maker. Such a market maker would not only overcome the problem of low liquidity by automatically accepting new orders but would also eliminate arbitrage opportunity trading by assuring market efficiency at any time (Hanson 2003).

3.4 A Laboratory Experiment on Group Technique Comparison

In addition to the TechForX field experiment, a laboratory experiment was conducted to compare prediction markets to three other group techniques on a quantitative judgment task: traditional face-to-face meetings, nominal groups, and the Delphi method. This experiment was conducted at the Wharton Business School at the University of Pennsylvania in Philadelphia, USA. The research was approved by the Institutional Review Board (IRB)[17] at the University of Pennsylvania, which requires empirical research involving human subjects to meet the ethical principles set forth in the Belmont Report.[18] Before being approved to conduct the experiment, it was required to pass the courses of CITI Protection of Human Subjects Research Training.[19]

3.4.1 Study design

Participants (n=227) were randomly assigned to 44 heterogeneous groups – 11 FTF, 11 NGT, 11 Delphi and 11 prediction markets.[20] Group size was determined by the number of students showing up in each session. Most groups (42 of 44) consisted of 4 to 6 subjects. One

[17] http://www.upenn.edu/regulatoryaffairs//index.php

[18] http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm

[19] http://www.citiprogram.org/

[20] The 11 prediction market groups were further divided into two treatments. In 5 groups, traders immediately started trading (PM).In 6 groups, traders independently generated their individual estimates before participating in the market (E-PM).

NGT group contained 3 subjects, one prediction markets group consisted of 7 subjects. Table 5 provides an overview of group size and the number of participants per method.

Table 5: Group size and number of participants per method

Group size	FTF	NGT	Delphi	Prediction markets	Total
3	-	1	-	-	**1**
4	6	4	1	-	**11**
5	2	1	5	5	**13**
6	3	5	5	5	**18**
7	-	-	-	1	**1**
No. of groups	**11**	**11**	**11**	**11**	**44**
No. of participants	**52**	**54**	**59**	**62**	**227**
Average group size	**4.7**	**4.9**	**5.4**	**5.6**	**5.2**

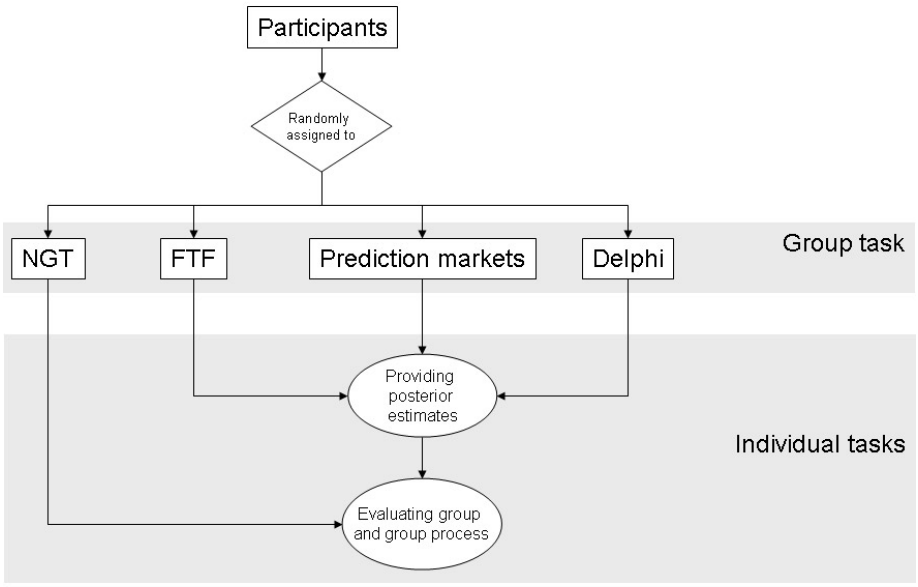
The experiment can be divided into three steps. First, participants entered a group interaction process to generate answers for 10 questions. Then, participants of meetings, Delphi and prediction markets worked independently to generate final individual estimates on the same 10 questions.[21] Finally, participants were asked to evaluate the group and the group process as a whole. The study design is illustrated in Figure 5.

3.4.2 Task type

The goal was to come up with answers for a quantitative judgment task, which consisted of 10 factual questions that required percentage estimates. These questions had correct solutions, although their correctness could not be verified at the time of the judgment. Thus, all group members could be expected to have some uncertainty about the answers. Decisive for the group to perform well was its ability to efficiently aggregate the relevant information of its members.

[21] In the case of NGT, the group results had to be manually calculated and, thus, were not known at the end of the session. Therefore, no posterior estimates could be obtained from participants in these groups.

Figure 5: Study Design – Laboratory Experiment



The question design was influenced by two limitations of the prediction market software. First, although, technically, prediction markets can be used to predict large numbers, it would have been necessary to scale the markets, which might have been harder to understand for participants. Second, for market maker mechanisms it is necessary to provide a starting price that should reasonably be chosen greater than 0. However, announcing starting prices provides anchors to traders.

To circumvent these problems, the quantitative judgment task consisted of 10 Almanac questions that required percentage estimates. That way, no scaling was necessary since market prices would always be between 0 and 100. Furthermore, all questions were designed similar by providing an anchor in the phrasing of the question which was used as the starting price in the markets. The questions and answers were taken from the *World Almanac and*

Book of Facts 2008[22] and were selected based on independent ratings from four coders. Table 6 lists the questions used in this study as well as the correct answers.

Table 6: Quantitative judgment task: ten Almanac questions

Question No. (as referred to in this work)[23]	**Question**	**Correct answer**
1	In 1960, the percentage of the total US population that was aged under 5 years was 11.3%. What was the percentage in 2000?	6.8
2	In 2006, the US share of worldwide Internet users was 17.3%. What share of worldwide Internet users did China have in 2006?	10.8
3	In 1950, the US share of the world motor vehicle production was 75.7%. What was the US share in 2006?	23.1
4	In 1970, 6.9% of the gross domestic product (GDP) was spent on health expenditures. What was the percentage spent in 2003?	15.2
5	In 2000, 13.3% of the world population lived in Africa. What was the percentage living in Australia (incl. Oceania) in 2000?	0.3
6	In 1980, 17.0% of the US population (25 years and over) had completed 4 years of college or more. What was the percentage in 2006?	28.0
7	In 1970, the percentage of all US children (under 18 years) who lived with two parents was 85.2%. What was the percentage in 2005?	67.3
8	In 1900, the percentage of the total US population that was aged 65 and over was 4.1 %. What was the percentage in 2000?	12.4
9	In 1994, the percentage of all US households that had a personal computer was 24.1%. What was the percentage in 2002?	60.0
10	In 1974, 4.0% of US children between 6 and 11 years old were overweight. What was the percentage in 2004?	18.8

Almanac questions have been used in a variety of studies to analyze group decision-making (e.g. Rowe & Wright 1999) and are representative of many problems in the "real world" – for example, "In 2008, our firm's market share is 4.2%. What do you predict it will be in 2010?"

It was emphasized in the written instructions and by the incentive mechanism that the purpose of the task was to estimate the answers to the questions. Answers were not revealed until completion of the study.

[22] See Janssen (2007)

[23] Note that the question numeration differs from the original order in the questionnaires (cf. Appendix M.2). As described in Section 6.3, for the analyses in this work, questions were ordered by ascending prior confidence of participants.

3.4.3 Participants

All participants were students at the University of Pennsylvania and were recruited by the Wharton Behavioral Lab. Each participant was paid a $10 show-up fee.

3.4.3 Incentive mechanism

In addition to the show-up fee, participants were remunerated depending on their group or individual performance. On average, $20 was distributed among each group. Thereof, on average, a share of $15 was distributed depending on a group's performance. In FTF, NGT, and Delphi, $50 were equally distributed among participants of the two most accurate groups, the next two most accurate groups received $25, and the next $15. To meet the traditional pay-off mechanism of prediction markets, traders were paid based on individual performance, i.e. $6 for the best performing trader in each group, $5 for the second best, and $4 for the third best.

For providing the final individual estimates, the two most accurate participants over all 10 questions in each group were remunerated (1^{st}: $3; 2^{nd}: $2). Before group interaction was finished, participants did not know that they would have to provide final individual estimates and had no information about potential additional pay-offs.

3.4.4 Materials and procedures

Data was collected in the Wharton Behavioral Lab. All sessions lasted for one hour. FTF and NGT were conducted in a small meeting room (see Figure 6). Delphi and prediction market groups worked in a computer room on individual workstations, divided by partition walls to prevent personal communication (see Figure 7).

Figure 6: Behavioral Lab – Meeting room

Figure 7: Behavioral Lab – Computer room

Each participant received general instructions explaining the relevant group technique as well as the respective pay-off mechanism. When the process started, forms were handed out to each participant for making personal notes and analyses. During the whole process, whenever individuals or groups were asked to reveal estimates, they had to state their confidence in these estimates on a seven point numerical scale (1: not at all confident; 7: extremely confident). After the group interaction, participants of all group techniques received a post-task form to reveal their final individual estimates, again along with confidence ratings. In addition, the post-task form was used to obtain participants' perceptions of the group and the group process as a whole, again on a seven point numerical scale. Examples of the used materials can be found in Appendix M.2.

3.4.4.1 Face-to-face meetings

Group members were seated around a table to discuss the problem with the goal of reaching a group estimate. Each group received one group questionnaire for their final group estimates. In addition, each group member received an individual questionnaire to note the achieved group estimates.

3.4.4.2 Nominal groups

First, group members received an individual questionnaire and were spread over three tables to work independently. Then, group members were seated around one table to discuss the

problem; each group received one group questionnaire for their final group estimates. Thereby, group members could independently decide whether they disclosed their prior estimates to the group or not. In addition, each group member received an individual questionnaire to note the achieved group estimates. Finally, group members were again spread out over three tables to provide their final individual estimates on the post-task form. The group results were calculated as the median of the final individual estimates.

3.4.4.3 Delphi

Before logging into the system, participants watched a video tutorial of how to use the Delphi software.[24] To conduct our research in a realistic environment that can be adopted by practitioners, the free Delphi software, available at forecastingprinciples.com, was used. In each round, participants provided their individual numerical estimates as well as lower and upper confidence bounds. Furthermore, participants had the possibility to reveal comments. After everyone responded, the results of the first round were summarized and reported as feedback to participants. This included participants' comments, their individual estimates along with confidence intervals as well as the statistical mean, median, and the standard deviation of these estimates. Then, participants provided their final individual estimates in a second round. For the analysis, the median of the individual estimates of the second round was used to calculate the group results.

3.4.4.4 Prediction markets

The (partly) free software solution of inklingmarkets.com was used since it was specifically designed to make participation as easy as possible for non-experienced traders and is easily available for practitioners. In addition, since the markets would have only a small number of traders, it was necessary to rely on a system using a market maker. Inkling's market maker provided sufficient liquidity, which was expected to increase trading activity. For further

[24] This tutorial can be downloaded at http://andreas-graefe.org/data/Delphi_Tutorial.ppsx.

discussion of the software see Christiansen (2007), who conducted field experiments using Inkling.

Before logging into the system, traders watched a video tutorial of how to use the prediction market software.[25] Each participant had an initial deposit value of $30,000 of play-money. At market closure, the last traded contract prices were interpreted as the final group results.

[25] This tutorial can be downloaded at http://andreas-graefe.org/data/pm_tutorial.pps.

Chapter 4

Validity of Prediction Markets for Long-term Forecasting

So far the track record of prediction markets is based on forecasting events in the near future. The literature provides little evidence on whether prediction markets are applicable for long-term problems whose outcome cannot be judged at the time the prediction is being made. Yet, decision-making must not be limited to the near future but has to consider the long-term.

This chapter reports on findings from the TechForX field experiment, described in Section 3.3, in which two prediction markets – one expert market and one student market – were implemented in parallel to a Delphi study. The goal of this forecasting exercise was to predict a classic judgment task (cf. Section 3.2): long-term technological trends whose outcomes could not be judged upon their occurrence.

4.1 Related Work

Traditionally, prediction markets are used to forecast events in the near future whose outcome will eventually be known and can clearly be judged. Such unambiguous outcomes make it easy to define a clear pay-off function for the remuneration of participants. This simplifies both

implementation and performance analysis of prediction markets since the accuracy of results can be judged ex post on the basis of the de facto occurrence (or non-occurrence) of the underlying event.

In contrast, little is known about the applicability and performance of prediction markets when it comes to the long-term. Results from an analysis of the *Foresight Exchange (FX)*[26], a prediction market aiming at assessing long-term developments, suggested that prediction markets might perform well for longer forecasting horizons (Pennock et al. 2001a). Launched in 1994, the FX is an online play-money prediction market that aims at forecasting events in the far future. For example, at the time of writing, one claim ('Lunar Excursion Tourism by 2025') predicted a 20% chance that a private tourist will land on the surface of the moon by January 1, 2025. Thus, this claim cannot be judged before January 1, 2025. For 161 contracts that referred to 'yes' or 'no' questions, Pennock et al. (2001a) recorded FX forecasts thirty days before the respective outcome was known. They found that the FX forecasts strongly correlated with outcome frequencies. However, the forecasting horizon of thirty days does not allow for drawing conclusions on the long-term forecasting performance of prediction markets. Furthermore, the analysis lacked a comparison to a benchmark forecast.

The question remains whether markets can generate reliable results for classic judgment tasks, like events whose outcomes cannot be judged at the time the prediction is being made. The lack of studies goes back to two reasons.

- Such outcomes preclude the definition of a clear pay-off function which is a basic requirement for the proper functioning of the market mechanism and provides incentives for participation.
- It is impossible to validate the accuracy of the results before the actual outcome can be judged. It just cannot be known today.

[26] http://www.ideosphere.com/

A way to solve this problem is to use external benchmarks to verify prediction market outcomes for judgment tasks. For example, Soukhoroukova (2007) used prediction markets to evaluate new product ideas as well as new product concepts. These are classic judgment tasks as they do not have demonstrably correct solutions but ask participants to come up with a consensus decision. For evaluating new product concepts, she found that the markets provided results that were consistent with participants' self-explicated expectations. Furthermore, Soukhoroukova (2007, p.90) found that the results yielded "an acceptable consistency with the results of traditional methods in product planning", like conjoint analyses. However, in the case of evaluating new product ideas, the market results did not reflect the assessments of an expert panel. In a similar study, Dahan et al. (2007) also used prediction markets to assess new product concepts. They found high consistency of market results with the results from an independent survey study. In sum, these findings suggest that prediction markets are applicable and can provide reasonable results for judgment tasks.

4.2 Hypotheses

In the TechForX experiment, a similar procedure was adopted to analyze the validity of prediction market results for a judgment task. The forecasts derived from two markets – an expert and a student market – were compared to the results of a well-established approach in judgmental long-term forecasting: the Delphi method. For a description of the experimental setting see Section 3.3.

Needless to say, the Delphi results cannot be considered as the 'ultimate truth'. However, Delphi is a well-established method for judgmental long-term forecasting. Thus, a similarity of market and Delphi results would provide support for the applicability of prediction markets for long-term problems.

As reported above, earlier studies showed that prediction market prices conformed to results derived from established methods for assessing judgment tasks. Although these studies

focused on the field of market research, it was expected that the findings would be similar for the field of long-term forecasting. Thus, it was hypothesized that the results of both markets would strongly correlate with the Delphi results. Thereby, the assumption that the student market would lead to similar results than the Delphi study might be surprising. After all, participants in the Delphi study were experts on the respective topic, whereas students could be expected to possess less expertise. Yet, as described in Section 1.2.3, research has shown that experts have little value in forecasting change. Thus, it was expected that the student market would perform similarly.

4.3 Results

The data used in this analysis can be found online.[27] For the following analysis, the final market prices were normalized to the portfolio price of 100. Delphi and market results were compared by calculating Spearman's rho correlation coefficients. The results are shown in Table 7. Both the student and the expert market correlated with the Delphi results with a Spearman's rho of 0.6. These correlations were significant.[28]

Table 7: Correlation of TechForX and EPIS Delphi results

	Students	Experts
Spearman's rho correlation coefficient	0.6	0.6

Figure 8 illustrates the correlation between the Delphi results and the results of both prediction markets. Note that, to some extent, the markets' deviation from the Delphi results might be explained by the well-known favorite-longshot bias, which has been observed in racetrack betting (Thaler & Ziemba 1988) as well as prediction markets (Wolfers & Zitzewitz 2004). According to this bias, individuals tend to overestimate low – and underestimate high –

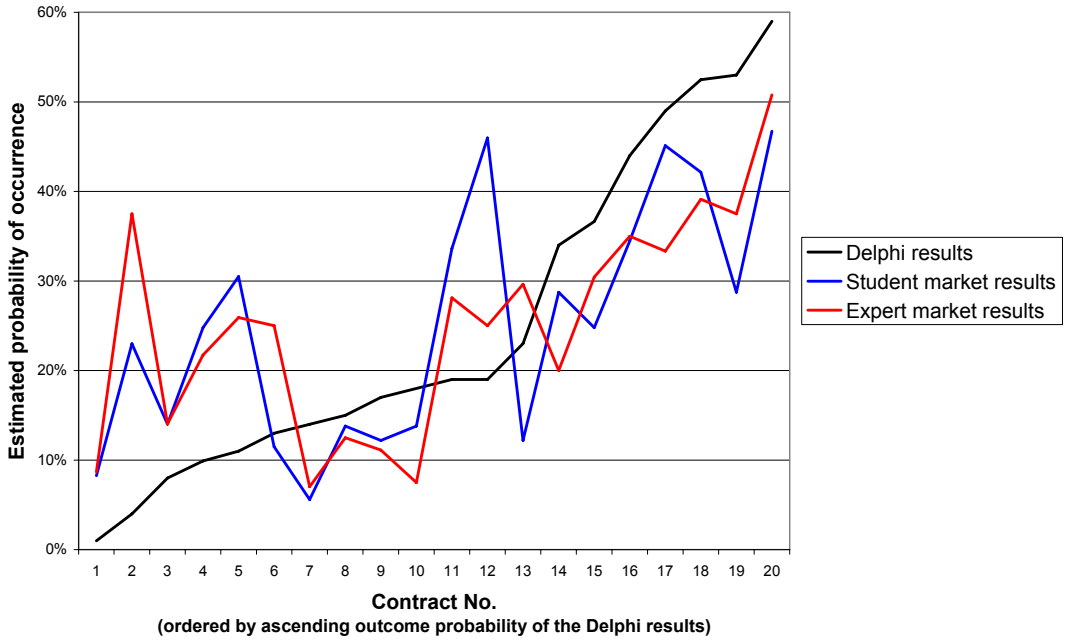
[27] http://spreadsheets.google.com/pub?key=pr1ZdfEZ874kIU-F05RnsIg

[28] Since the data was not normally distributed, the non-parametric Spearman's rho correlation coefficient was used. For both markets, Spearman's rho was statistically significant ($p<0.01$).

probabilities, a form of miscalibration that is also known as over-extremity pattern in psychology (Koehler et al. 2002).

It appears as if both groups, students and experts, systematically overvalued contracts for which the Delphi predicted probabilities of less or equal to 11%. At the same time, both groups systematically undervalued probabilities from 35% and above.

Figure 8: Correlation of Results (Delphi, Students, and Experts)



4.4 Discussion

The results suggest that prediction markets can provide reliable results for long-term forecasting problems. Interestingly, the resulting correlation coefficients of 0.6 for each market resemble the results of Soukhoroukova (2007). In comparing her markets to results from a conjoint analysis on a judgment task (i.e. the assessment of new product concepts), she derived a comparable (but slightly higher) correlation of 0.7.

Compared to short-term prediction markets these correlations might seem poor. Such markets measure ex post correlation based on the actual outcome of an event and often reach much higher correlations. However, one has to keep in mind that the TechForX markets aimed at predicting long-term technological trends which are characterized by high uncertainty about the actual outcome.

However, the TechForX markets required an external benchmark (i.e. the results of the Delphi study) to define the pay-off function for remunerating participants. It is still an open question how prediction markets might be implemented as a stand-alone method for long-term problems with uncertain outcomes. Yet, this should not prevent from implementing such markets. As shown by Armstrong (2001) in a meta-analysis of 30 empirical studies, combining forecasts from different methods and different sources generally improves forecasting accuracy. On average, Armstrong found a reduction in ex ante errors for equally weighted combined forecasts of 12.5%. He concluded that combining forecasts is especially helpful in situations involving high uncertainty and when one aims at avoiding large errors. Given these findings, prediction markets should be regarded as a complement to existing forecasting methods that can provide useful forecasts for long-term problems.

4.5 Summary

Within a European Foresight Project that aimed to assess long-term developments in the creative content sector, two prediction markets were implemented as an experimental add-on to a Delphi study. It was shown that both markets generated results that strongly correlated to the outcome of the Delphi study, whereby, to some extent, deviations from the Delphi results might be explained by the favorite long-shot bias. Based on a small sample, the results suggested that prediction markets do not have to be limited to events in the near future. Given the consistence with findings from the field of market research, prediction markets appear to be applicable and to provide reliable results also for classic judgment tasks, like long-term forecasting problems.

Chapter 5

The Value of Experts in Prediction Markets

When making decisions, people traditionally rely on experts who are believed to possess superior knowledge. In contrast, prediction markets are built on the idea of the 'wisdom of crowds' (Surowiecki 2004): they do not aim at separating experts from non-experts but are usually open for everyone to participate. However, as has become evident with the dismissal of the Policy Analysis Market (see Footnote 4), the idea of involving amateur knowledge into decision-making appeared to be a barrier for prediction markets (Hanson 2006). The question arises whether it can be valuable to limit market participation to experts.

This chapter reports on additional data from the TechForX experiment described in Section 3.3. In particular, it will be analyzed whether there is support for the assumption that experts possess more knowledge and, thus, might produce more reliable results than laymen.

5.1 Related work

Although findings from empirical research dispute the value of experts when predicting change (see Section 1.2.3), decision-makers favor expert advice. The reason might be that

expert judgment increases reliability and credibility of decisions or, at least, enables decision-makers to share responsibility.

Delphi is traditionally an expert-based approach and facilitators aim at recruiting participants with expertise. Thereby, one often aims at identifying the 'real' experts within the group of participants. Then, the judgments of these experts would be given higher weightings. A standard approach for identifying experts within participants is obtaining self-rated confidence scores. However, this implicates certain limitations due to evidence in the literature showing that people are typically overconfident (Einhorn & Hogarth 1978). Although Best (1974) and Rowe and Wright (1996) appeared to find that self-ratings as weightings can be valuable, others found no relationship between self-rated and actual expertise or accuracy (e.g. Brockhoff 1975, Dietz 1987, Larreche & Moinpour 1983, Sniezek & Henry 1990).

In general, weighting of judgments is a delicate issue. For methods like Delphi, assessing different weightings according to experts' confidence on various judgments would require complex mechanisms (Larreche & Moinpour 1983). Since there is no empirical evidence on how to appropriately assign weightings, Murphy et al. (1998) as well as Rowe and Wright (2001) suggested to generally circumvent weighting of judgments.

Prediction markets may improve on traditional approaches by implying a built-in weighting mechanism that automatically adjusts each individual judgment according to the (subconscious) confidence of a participant: the trading volume. The more confident a participant is, the more money he should be willing to invest. Accordingly, the higher his respective trading volume and, thus, the higher his influence on the market results. Vice versa, in case he is less confident, he should place orders with a lower trading volume and, thus, exert less influence on the results. Since participants are not directly asked to provide self-rated confidence estimates, they might not be aware of revealing their confidence

through the process of trading. In turn, this might lead to more objective confidence estimates.

Analyses of trading volume are common in financial markets research to measure confidence of market participants and a number of theoretical models suggest that higher confidence increases trading volume (e.g. Caballe & Sakovics 2003, Odean 1998). In an empirical analysis, Glaser and Weber (2007) confirmed recent research, showing that people who see themselves above average trade higher volumes. In addition, they found that higher trading volumes were not related to overconfidence.

5.2 Hypotheses

As in Chapter 4, the Delphi results were regarded as a benchmark for the validity of results. Given evidence that experts are of little value in forecasting, it was expected that, on average, the results of the expert and the student group would deviate about equally from the Delphi results. This would also conform to the findings from Chapter 4, where it was shown that both the expert and the student market correlated equally with the Delphi results.

H_1: On average, the results of both groups will deviate equally from the Delphi results.

But even if both groups would generate similar forecasts, experts should still have superior knowledge, which should result in higher levels of confidence. It was hypothesized that:

H_2: Experts will be more confident than students.

As listed in Table 3, the selected theses addressed different topics and technologies. Accordingly, the knowledge of a group could be expected to vary depending on the topic. On some topics, a group might have been better informed and, thus, should have been more confident than on others. This led to:

H_3: Confidence of a group will vary depending on the theses.

As described in Section 3.3.2, the task was to assess the realization time of each thesis within four time horizons. Given the design of long-term Delphi studies, it was expected that the four time horizons varied in terms of difficulty.

The Delphi theses were designed to meet the balance of being realistic but forward-looking and not too easy to predict. A typical thesis is usually not on the brink of being realized. On the other hand, it should also not be completely unrealistic that it might be realized at all. As a result, it is common for outcomes of long-term Delphi studies that most people would judge a theses' realization as most likely between the next 5 to 15 years. Thus, the time horizon 'In 2010-2020' was considered as an easy-to-assess environment. Furthermore, it seems feasible to assess whether a thesis will come true within the next three years, for which reason the time horizon 'Up to 2010' was classified as moderate-to-assess. By comparison, it appears harder to estimate whether a thesis will be realized in the far future ('Later') or not at all ('Never'). Therefore, these time horizons were classified as hard-to-assess.[29]

As stated above, experts are generally believed to possess superior knowledge. If so, such differences compared to students should become clear for forecasting environments that are hard-to-assess. It was expected that:

H_4: Experts will be more confident than students for hard-to-assess time horizons.

[29] This classification is subject to some subjectivity. However, as shown in Appendix M.1.3, classifying the time horizons based on the Delphi results leads to the same categorization.

5.3 Results

The data analyzed in this chapter can be found online.[30]

5.3.1 Relative validity of students' and experts' results

It was assumed that, compared to the Delphi results, on average, experts' forecasts would not differ substantially from those provided by the students (H_1). The mean absolute percentage error (MAPE) was used as a measure for the deviation of the results of each market from the Delphi results.[31] The smaller the MAPE, the smaller was the deviation from the Delphi results, and vice versa. For both groups, the APEs were calculated for each of the 20 contracts. Although the MAPE in the expert group (1.09) was smaller than in the student group (1.24), differences were small.[32] Hypothesis H_1 was corroborated.

5.3.2 Confidence of experts and students compared

In the following, the order volume was interpreted as a measure of participants' confidence. It is defined as the price of an order multiplied by the number of shares.

5.3.2.1 Group confidence at a glance

It was assumed that experts possess superior knowledge and, thus, would be generally more confident than students (H_2). For both groups, Table 8 shows the mean order volumes (MOV), as well as the respective 95% confidence intervals, over all 20 contracts. While students were about four times more active, experts traded higher volumes than students. This indicates that experts were more confident in their judgments. These differences were

[30] http://spreadsheets.google.com/pub?key=pr1ZdfEZ874kJrmlGDHLvJA

[31] The definition of the MAPE is provided in the Technical Appendix.

[32] For neither group, the APEs were normally distributed. Thus, a Mann-Whitney U-test was conducted to compare the APEs of the student and the expert markets. The test showed no significant differences between both groups.

significant.[33] This can also be seen by looking at the 95% confidence intervals, which show little overlapping. Hypothesis H_2 was corroborated.

Table 8: Group confidence at a glance

	No. of orders	Mean order volume	95% confidence interval	
			Lower limit	Upper limit
Students	1,672	1,961	1,869	2,051
Experts	403	2,174	1,984	2,364

5.3.2.2 Confidence between theses

It was expected that a group's knowledge would vary depending on the theses and, thus, confidence for different theses should vary, too (H_3). For each thesis, the MOVs (and the 95% confidence intervals) for both groups were calculated. Table 9 shows the results. For the students, differences in the MOVs between theses were small.[34] The interval between maximum (thesis 2) and minimum (thesis 3) MOV for all theses was 263 (2,102−1,839). Thus, hypothesis H_3 was rejected for the student group: students were about equally confident for all theses.

By comparison, in the expert group, variations of the MOVs between theses were significant.[35] Here, the MOV ranged from a minimum of 1,582 (thesis 1) to a maximum of

[33] A one-tailed t-test was conducted to compare the MOVs of students and experts. The test showed statistically significant differences between both groups ($p=0.02$). However, the results of the t-test have to be interpreted carefully since the order volume data was not normally distributed. That said, a number of authors argue that the t-test is robust to violations of the normality assumption (Glass et al. 1972, Rider 1929). In addition, the U-test showed similar results.

[34] A Kruskal-Wallis-test was conducted to compare the mean order volumes of a group between theses. For the students, the test showed no statistically significant differences.

[35] In the case of the expert group, the Kruskal-Wallis-test showed statistically significant differences ($p<0.001$). Yet, due to a large number of tied ranks in the data, the results should be interpreted carefully.

2,723 (thesis 2). Thus, hypothesis H_3 was corroborated for the expert group: experts' confidence varied depending on the theses.

Table 9: Group confidence between theses

Thesis	Students			Experts		
	Mean order volume	95 % confidence interval		Mean order volume	95 % confidence interval	
		Lower limit	Upper limit		Lower limit	Upper limit
1**	2,010	1,789	2,232	1,582	1,320	1,845
2*	2,102	1,864	2,340	2,723	2,057	3,390
3**	1,839	1,650	2,029	2,516	1,959	3,073
4	1,965	1,767	2,163	2,082	1,688	2,468
5*	1,916	1,727	2,104	2,364	2,000	2,729

In addition, Table 9 shows how theses were judged *between both groups*. The MOVs for the theses 1, 2, 3, and 5 differed significantly.[36] Merely for thesis 4, differences were small (i.e. experts and students were about equally confident). This can also be seen by looking at the confidence intervals. The interval for the student group fell within the interval for the expert group. Given the results from above, one might assume that experts were generally more confident. This was not the case. Although for three out of four significantly different theses (i.e. theses 2, 3, and 5) the experts' MOV was higher, it was lower for thesis 1.

5.3.2.3 Confidence along time horizons

It was expected that experts would be more confident for hard-to-assess forecasting environments (H_4). For each time horizon, Table 10 shows the MOVs of both groups. For the easy-to-assess time horizon '2011-2020', students were more confident than experts, although differences were small (both 95% confidence intervals showed large overlapping).[37]

[36] To compare mean order volumes for students and experts, two-tailed t-tests were conducted for each thesis.

[37] A two-tailed t-test, conducted to compare order volumes of students and experts, showed no statistical significant differences.

Experts were more confident for the three remaining time horizons, with significant differences for both hard-to-assess environments.[38] Hypothesis H_4 was corroborated.

Table 10: Group confidence along time horizons

		Students			Experts		
			95 % confidence interval			95 % confidence interval	
Time Horizon	Difficulty to assess	Mean order volume	Lower limit	Upper limit	Mean order volume	Lower limit	Upper limit
Up to 2010	Moderate	1,371	1,246	1,496	1,642	1,361	1,923
2011-2020	Easy	3,032	2,793	3,272	2,859	2,362	3,356
Later*	Hard	2,024	1,849	2,198	2,419	2,015	2,823
Never**	Hard	1,454	1,322	1,587	1,903	1,581	2,225

The results were similar to the comparison of group confidence along theses: experts were not generally more confident than students. For the easy-to-assess environment, whose high probability of occurrence might have been predictable without specific knowledge, students were equally confident than experts. This indicates that experts did not exhibit overconfidence. Rather, the results suggest that experts possessed more specific knowledge: they were more confident for moderate- and hard-to-assess forecasting environments.

5.4 Discussion

It was found that, on average, the results of the expert and the student market deviated equally from the Delphi results. This is not surprising. As described in Section 1.2.3, earlier research showed that experts have little value in forecasting change. However, an analysis of the order volumes, which were interpreted as a measure of confidence, revealed differences in the trading behavior of experts and students.

---

[38] Two-tailed t-tests were conducted to compare order volumes of students and experts for each time horizon. Note that for testing hypothesis H_4, a one-tailed t-test would have been sufficient. Although not expected, the results of the *one*-tailed t-test showed that experts were significantly more confident also for the moderate-to-assess time horizon 'Up to 2010' ($p<0.05$).

1. Overall, experts were more confident than students. This was expected as experts can be assumed to possess more specific knowledge than students.
2. Due to the variety of topics, it was expected that a group's knowledge – and, thus, confidence – would differ along theses.
    - The order volumes in the student market did not differ along the five theses. This indicates that, regardless the topic, students were equally confident in their assessments. A reason might be that students did not possess specific information depending on the topic. It appears as if they simply 'guessed' by revealing information through trading.
    - By comparison, experts' confidence varied significantly between the five theses. This lets assume that experts revealed their beliefs based specifically on what they thought they would know. For theses for which they believed to possess relevant information, they might have been more confident. Vice versa, in situations where they believed to know less, they appeared to be less confident.
3. Due to differences in expertise, it was expected that experts would be more confident than students for hard-to-assess forecasting environments. The results corroborated this hypothesis. In addition, experts were also more confident for the moderate-to-assess forecasting environment.
4. Given the results, one might assume that experts were generally more confident, which could be interpreted as overconfidence. However, two findings indicate that this was not the case:
    - For the easy-to-assess environment, students were more confident, although differences were small.
    - Comparing confidence between groups for each thesis revealed that experts were not generally more confident. Although they were more confident for three theses (no. 2, 3, and 5), they were also less confident for thesis 1. For thesis 2, experts' and students' confidence did not differ.

These results conform to findings from the financial markets literature. In an empirical study, Glaser and Weber (2007) compared results derived from calibration questions from an online questionnaire to trading volumes of 215 individuals. They found that people who think they are above average trade higher volumes, whereas the higher volumes were not related to overconfidence.

Although both markets provided similar results, the findings suggest that experts revealed information well-considered, based specifically on what they thought they would know. By comparison, it appears as if students 'just traded'.

Such information derived from analyses of trading behavior can be highly valuable for decision-makers. It discloses participants' confidence and, thus, sheds light on their expertise. When assessing long-term developments involving high uncertainties, decision-makers should be more willing to rely on judgments derived from an informed group as this adds credibility and reliability.

In incorporating a built-in measure of participants' confidence in their own beliefs (i.e. the order volume), prediction markets can improve on traditional approaches of information aggregation. By comparison, Delphi has no implicit mechanism to measure confidence. Instead, often self-rated confidence scores are obtained by asking participants to specify their level of expertise before answering a particular thesis. Yet, earlier research reported limitations of self-rated confidence scores. Confidence levels obtained from prediction markets appear to be less vulnerable to wrong self-assessment: Since participants are not directly asked about their level of expertise, they may not be aware of revealing their confidence as this happens subconsciously through the process of trading. Thus, analysis of traded order volume might lead to more objective confidence estimates.

5.5 Summary

In comparing results and trading behavior of an expert and a student market from the TechForX field experiment (cf. Section 3.3), this chapter analyzed the value of expert-based prediction markets for long-term forecasting problems. While the results of both markets did not differ compared to the benchmark Delphi study, the analyses identified differences in the trading behavior of both groups. Experts appeared to reveal their beliefs well-considered based on what they think they know. By comparison, the trading behavior of students led to the assumption that students did not possess specific knowledge but 'just traded'.

Analyses of participants' trading behavior provide new opportunities to extract relevant information for decision-makers from prediction markets as they give insights into the reliability and credibility of forecasts, especially in situations involving high uncertainty. The present study should motivate further research on analyses of trading behavior, a field that appears valuable for improving the usefulness of prediction markets for decision-makers.

Chapter 6

Relative Accuracy of Prediction Markets on a Quantitative Judgment Task

A variety of approaches can help to elicit information from groups. While organizations most commonly rely on unstructured face-to-face meetings, it is difficult to find evidence to support the use of this strategy. Evidence from the literature suggests that structured approaches like nominal groups or Delphi allow for more accurate forecasts than traditional meetings. However, little is known about the relative performance of prediction markets compared to meetings as well as to other structured group techniques for aggregating the knowledge in groups.

To address this deficit, a laboratory experiment was conducted to compare unstructured meetings, nominal groups, the Delphi method and prediction markets. This chapter reports on the relative accuracy of the four group techniques on a quantitative judgment task. For a description of the study design see Section 3.4.

6.1 Related Work

As described in Section 3.1.1, meetings can be subject to various types of biases. For example, group members often fail to disclose information due to an announcement by

others which made them uncertain about their own beliefs. If individual opinions differ from the group opinion, group members tend to ignore their private information. Furthermore, particularly in hierarchical meetings, individuals may silence themselves because other group members impose social pressure on them. Although they might believe that they know better, they will not share their information because they are afraid of sanctions.

Structured group techniques are expected to limit these biases and prior research seems to support this assumption:

- In their meta-analysis, Rowe and Wright (1999) reported superior accuracy of Delphi over unstructured interaction by a score of five studies to one, with two ties. However, their analysis included two studies that did not refer to quantitative judgment tasks. Excluding these two studies led to an adjusted score of three studies to one, still with two ties.

- Similarly, in summarizing the literature, Woudenberg (1991) found a – albeit slight – superiority of Delphi over unstructured interaction.

- Furthermore, again for quantitative judgment tasks, Gustafson et al. (1973) found NGT to be more accurate than FTF, although Fischer (1981) found no differences.

- In case of prediction markets, to date there is no study that compared the approach to unstructured interaction on this type of tasks.

Evidence on the relative performance of NGT, Delphi, and prediction markets for quantitative tasks is scarce. While Gustafson et al. (1973) showed that NGT was more accurate than Delphi, two studies found no differences (Boje & Murnighan 1982, Fischer 1981). Again, for prediction markets, no work is known that compared the method to NGT

or Delphi. Thereby, arguments for their relative performance could be found in either direction (cf. Section 3.1.3.3). One might argue that prediction markets would be advantageous to Delphi since traders can continuously update their estimates and are not tied to a limited number of rounds. On the other hand, traders usually do not reveal reasons or comments for their estimates. Thus, other market participants cannot relate to *why* the group estimate has changed. Accordingly, the possibility of argumentation and reasoning might favor NGT and Delphi.

6.2 Hypotheses

Given the lack of evidence, there was no directed hypothesis on the relative accuracy of the three structured approaches NGT, Delphi, and prediction markets. Yet, it was expected that all three structured approaches would outperform FTF:

$$\text{NGT} = \text{Delphi} = \text{prediction markets} > \text{FTF}$$

The study design was sent to experts in the field who were asked to provide their prior estimates on the relative accuracy of the methods. To assure that the experts were familiar with all four methods, primarily people experienced with prediction markets were contacted. Eight experts[39] responded and the following ranking was derived from their priors: Delphi > prediction markets > NGT > FTF. Expecting that all three structured approaches would outperform FTF, the expert priors were consistent with the above hypothesis. In particular, none of the experts expected FTF to do better than any of the structured approaches. In addition, the audience of two conferences[40] was asked to reveal

[39] Thanks to Yiling Chen, Kesten C. Green, Robin Hanson, Stefan Luckner, Gene Rowe, Martin Spann, Gerrit H. Van Bruggen and Eric Zitzewitz.

[40] The study has been presented at the *28th International Symposium on Forecasting* in Nice, France (Graefe 2008b) and the *Third Workshop on Prediction Markets* in Chicago, USA (Graefe 2008c).

their priors. The results were similar. Both audiences expected FTF to perform worst. However, Delphi and prediction markets switched places: at both conferences, the audience expected prediction markets to perform best, Delphi second and NGT third.

6.3 Results

The data that has been analyzed in this chapter can be found online.[41] The median absolute percentage error (MdAPE) was used as an index of accuracy.[42,43] The results are shown in Table 11. For each question, the MdAPEs of the four methods as well as the MdAPEs over all ten questions are reported. With the lowest MdAPE of 0.21, Delphi was most accurate, followed by NGT (0.25). As expected, with an overall median error ratio of 0.30, FTF was least accurate. Prediction markets ranked third (0.27).

Table 11: MdAPEs of FTF, NGT, Delphi and prediction markets

(sample size = 11 forecasts per question)

Question	FTF	NGT	Delphi	Prediction markets
1	0.12	0.25	0.18	0.31
2	0.30	0.35	0.44	0.32
3	0.52	0.52	0.65	0.89
4	0.12	0.05	0.21	0.18
5	5.67	7.33	12.33	17.33
6	0.21	0.18	0.11	0.09
7	0.08	0.11	0.03	0.10
8	0.61	0.40	0.23	0.57
9	0.17	0.25	0.17	0.11
10	0.60	0.17	0.12	0.52
Overall (N=110)	**0.30**	**0.25**	**0.21**	**0.27**

[41] http://spreadsheets.google.com/pub?key=pr1ZdfEZ874nHtHGW_CT7dA

[42] For the definition of MdAPE see the Technical Appendix.

[43] The analysis was replicated using the error ratio as a measure of accuracy. The results were similar.

As expected, overall, all three structured approaches outperformed FTF. Yet, differences were small.[44] This led to speculations whether there might be situations in which FTF can perform better than structured methods, and vice versa.

In the following, participants' prior confidence was interpreted as a measure for the difficulty of questions. The higher prior confidence, the easier it should have been for participants to come up with accurate estimates, and vice versa. Confidence ratings were obtained from prior individual estimates made in the NGT and E-PM groups, i.e. before participants entered group interaction (cf. Appendix M.2.2).[45] For each question, the last column in Table 12 provides the means of the individual confidence ratings. Questions were ordered by ascending confidence.[46] For question 1, participants were least confident; for question 10, they were most confident. Participants' mean confidence was generally low, ranging from 2.79 for question 1 to 3.46 for question 10. For none of the questions, mean confidence reached the median of the 7-point Likert scale and half of the questions obtained a confidence rating lower than 3.

Along with confidence scores, Table 12 reports the error reductions of NGT, Delphi, and prediction markets compared to FTF. The error reduction rates were calculated as the difference between the MdAPEs of FTF and the respective structured method. In case of positive error reduction, the respective structured method did better than FTF. In case of negative error reduction, the respective structured method did worse than FTF.

---

[44] Since the data was not normally distributed, a Kruskal-Wallis-test was conducted to compare the overall mean absolute percentage errors (MAPE) across the four methods. The results showed no statistically significant differences over all 10 questions.

[45] For 8 of the 10 questions, 87 individual confidence ratings were obtained. For questions 1 and 9, one subject did not provide confidence ratings, leading to 86 observations for these questions.

[46] Similarly, questions in Table 11 were already ordered by ascending confidence. As can be seen by comparing Table 3 and the tables in Appendix M.2, the original order of the questions in the study questionnaire differed.

Table 12: Error reduction of NGT, Delphi, and prediction markets compared to FTF

Question (ranked from lowest to highest confidence)	MdAPE FTF	Error reduction to FTF			Mean prior confidence
		NGT	Delphi	Prediction markets	
1	0.12	-0.16*	-0.08	-0.21**	2.79
2	0.30	-0.05	-0.12	-0.03	2.82
3	0.52	0	-0.09	-0.23**	2.82
4	0.12	0.11	-0.12	-0.09	2.85
5	5.67	-0.33	-1.13	-1.78	2.95
6	0.21	0.04	0.14	0.16	3.16
7	0.08	-0.05	0.10	-0.03	3.16
8	0.61	0.15	0.31**	0.03	3.31
9	0.17	-0.09	0	0.08	3.31
10	0.60	0.36**	0.43**	0.05	3.46
Overall	**0.30**	**0.04**	**0.09**	**0.03**	**-**

The error reductions reveal that the relative performance of the structured approaches compared to FTF was quite similar. FTF tended to outperform the three structured approaches for questions that obtained rather low confidence scores (questions 1 to 5). On the contrary, the structured approaches (in particular Delphi and prediction markets) tended to be more accurate for questions that achieved higher confidence scores (questions 6 to 10). For some questions, in particular questions on the edge of the confidence distribution, these differences were significant.[47]

It appears as if participants' confidence in their prior individual estimates, and thus question difficulty, affected the accuracy of the methods. To further analyze this relationship, the quartiles of the mean confidence ratings were observed. Thereby, questions for which participants were least confident (C<2.83; lower quartile), were considered as hard-to-estimate. Questions for which participants were highly confident (C>3.27; upper quartile), were defined as easy. Questions in the interquartile range were defined as moderate.

[47] Two-tailed t-tests were conducted to compare the MAPEs of FTF and the respective structured approaches.

For each quartile, the performance of FTF was analyzed relative to each structured approach.[48] The results are shown in Table 13, again reported as the error reduction of the three structured approaches compared to FTF. For hard-to-estimate questions, all three structured approaches were inferior to FTF. For Delphi and prediction markets, these differences were significant. For moderate- and easy-to-estimate questions, all three structured approaches outperformed FTF. For easy-to-estimate questions, these differences were significant. The results suggest that method accuracy was affected by difficulty of question.

Table 13: Question difficulty and method accuracy

Quartile	Questions	Confidence	Difficulty	MdAPE FTF	Error reduction to FTF		
					NGT	Delphi	Prediction Markets
Lower	1, 2, 3	C ≤ 2.83	Hard	0.26	-0.06	-0.15*	-0.17**
Interquartile	4, 5, 6, 7	2.83 > C < 3.27	Moderate	0.23	0.07	0.05	0.09
Upper	8, 9, 10	C ≥ 3.27	Easy	0.49	0.24**	0.32**	0.25*

6.4 Discussion

As expected, overall, the three structured approaches outperformed FTF. Delphi performed best, followed by NGT and prediction markets. However, differences were small and FTF did not generally worst. In analyzing prior confidence estimates, it was found that the relative accuracy of FTF and structured methods appeared to be affected by the difficulty of questions.

6.4.1 Method accuracy vs. question difficulty

All three structured methods outperformed FTF for easy-to-estimate questions. For such questions, the mean of participants' prior confidence was high, which may indicate that the

[48] For each quartile, the MAPEs of FTF and the respective structured approach were compared. For the lower and the upper quartile, two-tailed t-tests were conducted. For the interquartile range, a Mann-Whitney U-test was used since the underlying data was not normally distributed.

majority of group members possessed relevant information. Then, a group will perform well if this dispersed information will be efficiently aggregated. In such situations, structured approaches should be advantageous in incorporating the information of group members as they are less vulnerable to group pressures. The results seemed to confirm this.

On the contrary, FTF tended to outperform Delphi and prediction markets for hard-to-estimate questions; these differences were significant. One can only speculate on the reasons.

Participants in this study had to solve a quantitative judgment task that required percentage estimates on factual questions. Such tasks have clear solutions and require groups only to aggregate factual knowledge. In contrast, other tasks like decision-making, negotiation, or conflict solving are more challenging as they involve not only 'facts' but also values, emotions, expectations, etc. For such tasks, one would expect group pressures to be much higher and, therefore, FTF to perform generally worse. In addition, the experiment setting favored meetings since it did not involve hierarchies. In a more realistic environment with hierarchies, one would, again, expect FTF to perform worse.

For a group to perform well, it should only aggregate relevant information from its members. If irrelevant or poor information is aggregated, decision accuracy might be harmed. For hard-to-predict questions, mean confidence was low. Accordingly, one could speculate that only few group members possessed relevant information whereas a majority might have been poorly informed. To perform well in such situations, a group has to filter poor information and identify (and follow) the judgment of its best member(s). In such situations, intensive information exchange and reasoning, facilitated through direct interaction (like in FTF or NGT), might be necessary for the group to be convinced by its best-informed members. In fact, research has shown that interacting groups have some ability to identify their best members and incorporate their judgments in the group estimate (Henry 1993, Henry 1995).

On the contrary, approaches that disallow direct interaction might fail as participants' possibilities to provide reasoning are limited. This became most evident in the case of prediction markets. Although participants could exchange information continuously through trading, they were unable to provide reasoning for their estimates. They could only buy and sell contracts, without indicating why they did so. If only few group members were well-informed, this mechanism might not have allowed them to convey their information to the group and, thus, to convince the group of their views.

Yet, prediction markets could be designed in a way that they are able to deal with problems for which only few group members possess relevant information. In particular, they could be enhanced by allowing traders to communicate with each other in order to reveal reasoning for their actions, for example, by adding additional means of communication like blogging or chatting functionalities. Although several prediction markets already adopted such solutions, to date, no research is known that examined whether this has an effect on accuracy.

6.4.2 Impact of incentives

As described in Section 3.4.3, monetary incentives were provided to participants of each group technique. These were based either on participants' individual performance (prediction markets) or on the performance of their group (meetings, NGT, and Delphi). Incentives were provided as they are expected to motivate participants to perform the task well and, thus, to reduce experimental error (Remus et al. 1998). As can be derived from the meta-analysis of Rowe and Wright (1999), offering incentives for studies of group technique comparison is common in the forecasting and decision-making literature.

One might question whether the different incentive schemes may have had an impact on the results. Performance-based incentives – which can be monetary, tangible, or social (in form of a user ranking) – are an integral feature of prediction markets. By comparison, in real-

world situations, one usually does not provide performance-based (monetary) incentives to participants in meetings, NGT, or Delphi. Thus, one might argue that these group techniques were favored by the experimental setting. In turn, one might expect better performance of prediction markets in real-world situations.

There is a large amount of research on the impact of incentives on decision-making or forecasting accuracy. Interestingly, these studies tend to show that monetary incentives do *not* have a positive impact on accuracy:

- For a quantitative task, consisting of Almanac questions, Henry and Sniezek (1993) found no impact of monetary incentives on accuracy.

- Equally, in reporting on a judgmental forecasting experiment on a time series task, Remus et al. (1998) concluded that there was no evidence that monetary incentives impacted forecasting accuracy.

- For prediction markets, Luckner and Weinhardt (2007) analyzed three different incentive schemes in the STOCCER FIFA World Cup 2006 market. Their results showed that performance-based incentives did not necessarily increase prediction accuracy. This conforms to findings from studies that found little or no difference in accuracy between play-money and real-money prediction markets (Rosenbloom & Notz 2006, Servan-Schreiber et al. 2004).

- Finally, in reviewing the literature on the effects of monetary incentives on performance in laboratory tasks, Bonner et al. (2000) derived similar findings: monetary incentives improved performance in only about half of the 131 experiments that were analyzed in their study. Furthermore, they found that the positive impact of incentives on accuracy decreases with increasing task complexity. They argued that, for more complex tasks (like the judgment tasks used in the

present work), participants' average skill level decreases, which makes it less likely that incentives improve performance. By comparison, incentives tend to have a positive impact on performance for easier problems like vigilance or detection tasks, memory tasks, or clerical tasks.

Given these findings, there is little reason to believe that the experimental setting in the present laboratory experiment favored meetings, NGT, and Delphi over prediction markets. Nonetheless, it might be worthwhile to conduct further empirical studies that are specifically designed to analyze this question.

6.5 Summary

In reporting on data from the field experiment described in Section 3.4, this chapter compared the relative accuracy of traditional FTF to three structured approaches (NGT, Delphi, and prediction markets) on a quantitative judgment task. This task consisted of ten factual questions that required percentage estimates.

Overall, the structured approaches outperformed FTF with Delphi performing best, NGT second, and prediction markets third, although differences were small. However, the relative accuracy of the methods appeared to be affected by the difficulty of the questions. For easy-to-estimate questions, all three structured approaches did significantly better than FTF. This is not surprising as structured approaches should outperform meetings in situations where the majority of group members have relevant information. For hard-to-estimate questions, FTF outperformed Delphi and prediction markets. A reason might be that well-informed Delphi – and particularly prediction market –participants were limited in their ability to provide reasoning for their judgments. Thus, they failed in exerting influence on their group and convincing fellow group members of the quality of their judgment.

Chapter 7

Perceptions of Prediction Markets

When making decisions based on a forecast, one inevitably has to deal with uncertainty. Since the outcome of the underlying event cannot be known at the time the forecast is being made, one might question whether potential accuracy is the appropriate criterion for measuring the effectiveness of a forecasting – or decision-making – method. This is particularly true in situations in which there is often no one correct solution or where the decisions affect the lives and behavior of decision-makers. As a result, the objective, analytical quality of a decision can become secondary. In such situations, the dominant criteria for choosing a method might be participants' perceptions of the decision-making process, for example how satisfied participants were with the process itself or its outcome.

In other words, particularly in situations where an objective measure of quality does not exist, the quality of a decision is often measured by the degree of acceptability. If group members or decision-makers feel dissatisfied with the process, its outcome may not be adopted – even if highly accurate. However, needless to say, a process that is highly satisfying for participants may not necessarily lead to accurate results.

Practical experience indicates that people have problems in understanding how prediction markets work. If so, one can assume that people are not satisfied with participation in a

market and, therefore, might not trust the results. As suggested in Chapter 1, this can be a barrier for the implementation of prediction markets. This chapter reports on the third step of the laboratory experiment described in Section 3.4. In particular, it will be analyzed how participants of meetings, nominal groups, Delphi and prediction markets perceived their group as well as the group process a whole.

7.1 Related Work

Given the large number of group decision making studies, little work has been done on analyzing participants perceptions of group processes. For quantitative judgment tasks, Boje and Murnighan (1982) analyzed participants' perceptions for Delphi, NGT, and people working alone. They found that NGT was rated most favorable in terms of effectiveness, satisfaction and freedom to participate, whereas Delphi was rated only slightly superior compared to working alone. Van de Ven and Delbecq (1974) compared participants' perceptions of nominal groups, Delphi, and unstructured meetings for an idea generation problem. They found that NGT participants expressed highest satisfaction with the process whereas differences between Delphi and meetings were small. For prediction markets, thus far, no study is known that analyzed how people perceive participation and how this might affect confidence in market outcomes.

7.2 Hypotheses

In general, personal interaction in groups can either lead to coherence, and thus high perceived satisfaction, or disagreement, resulting in frustrated group members. In the present study, personal interaction in FTF and NGT was conducted as non-hierarchical meetings. Thus, little group pressures were expected. In addition, findings from the literature suggest that people generally enjoy human interaction and the sense of working together.

Thus, it was hypothesized that personal interaction in FTF and NGT would lead to more favorable ratings compared to Delphi and prediction markets. Furthermore, due to less conformity pressures in NGT, it was expected that NGT would be rated more favorable than FTF; a result that had been confirmed earlier by Van de Ven and Delbecq (1974). In addition, since the majority of participants were expected to have been unfamiliar with the process of revealing one's opinion by trading stocks, it was expected that prediction markets would be rated least favorable, in particular in terms of difficulty.

7.3 Results

The data that has been analyzed in this chapter can be found online.[49] For each category, participants' ratings, revealed on a 7-point Likert scale from '1' (very low) to '7' (very high), were summarized using the statistical mean. Then, for each category, a ranking of the methods from 1 (ranked most favorable) to 4 (ranked least favorable) was computed.[50] For an example of the underlying questionnaire see Appendix M.2.4.

7.3.1 Ratings of the group

The results are shown in Table 14. NGT obtained most favorable ratings for *cooperation* and *disagreement* and ranked second for participants' *confidence* in the group answers. Prediction markets ranked second worst for *disagreement* and *confidence* and worst for *cooperation*. For both *cooperation* and *disagreement*, FTF ranked second. Although Delphi obtained the highest score for *disagreement*, its participants' were most confident in the outcomes.

[49] http://spreadsheets.google.com/pub?key=pr1ZdfEZ874nHtHGW_CT7dA

[50] For the categories *disagreement* and *difficulty,* favorable ranks were assigned for low ratings.

A reason for the high perceived disagreement in Delphi might be that Delphi participants cannot communicate directly with each other but reveal their estimates independently. Thus, agreement can only increase if incorporating feedback information after the first round leads to more cohesive results at the end of the process. However, note that disclosing disagreement can also be desirable: it alerts decision-makers to uncertainty.

Table 14: Participants' perceptions of their groups

	Mean of individual ratings				Rank			
	FTF (n=52)	NGT (n=53)	Delphi (n=46)	Prediction markets (n=60)	FTF	NGT	Delphi	Prediction markets
Cooperation	5.9	6.4	5.0	4.2	2	1	3	4
Disagreement	3.5	2.8	4.0	3.9	2	1	4	3
Confidence	4.1	4.6	4.7	4.3	4	2	1	3

FTF obtained the lowest score in terms of *confidence* in the group outcome, ranking even behind prediction markets. This is surprising since, according to the theory of groupthink (Janis 1972), one could expect that high levels of perceived cooperation and agreement would result in high levels of confidence in the group results. It seems as if FTF participants were aware of such group tendency effects and realized that probably not all available information was aggregated from group members.

7.3.2 Ratings of the group process

Table 15 shows the results. Rankings of the group process as a whole were identical over 4 out of 5 categories with NGT placed first, FTF second, Delphi third and prediction markets fourth. For *difficulty to participate*, FTF achieved an equally favorable score as NGT. For *freedom to participate*, FTF and Delphi swapped places. Given the favorable ratings for FTF on most other categories, it is interesting that Delphi achieved a higher score than FTF for *freedom to participate*. Again, the reason might be that FTF participants experienced group pressures that may have hindered them to fully reveal their information, which conforms to the low confidence scores for FTF. Overall, participants' perceptions of the prediction

market process were poor for all categories. Using a two-tailed t-test, the differences between prediction markets and Delphi, as the second worst rated technique, were statistically significant for *freedom to participate*, *satisfaction,* and *difficulty*.

In sum, the results conform to the hypotheses. NGT obtained most favorable ratings for 7 of the 8 categories. FTF were rated quite favorable, too; ranking second for five categories and first for one category. Finally, the results corroborate the hypothesis that prediction markets would be rated most unfavorable. For 6 categories, prediction markets obtained the worst ratings and were second worst for the two remaining categories.

Table 15: Participants' perceptions of their group process

	Mean of individual ratings				Rank			
	FTF	**NGT**	**Delphi**	**PM**	**FTF**	**NGT**	**Delphi**	**PM**
Freedom to participate*	5.6	6.1	5.8	5.3	3	1	2	4
Time well spent	5.2	5.6	4.9	4.6	2	1	3	4
Difficulty**	2.3	2.3	2.7	3.8	1	1	3	4
Overall satisfaction*	5.4	5.6	5.1	4.6	2	1	3	4
Effectiveness	5.0	5.3	4.3	4.2	2	1	3	4

7.4 Discussion

Examining participants' perceptions of the group and the group process as a whole revealed a clear preference for methods involving personal communication. This supports the assumption that people simply enjoy human interaction. Thus, personal interaction in non-hierarchical meetings can lead to coherence and, therefore, increase perceived satisfaction.

In contrast, Delphi and prediction markets were rated less favorable. In particular, prediction markets were rated worst on 6 categories and second worst on the remaining two. There might be at least two reasons: First, the lack of exchanging reasons for judgment may alienate traders. One could expect that convenience with the method increases by gaining insight into the opinions of fellow traders. Second, prediction markets are still a relatively new approach and the idea of revealing one's opinion by trading contracts may be

difficult to understand. Even though the experiment software was specifically designed to make participation as easy as possible for novices, participation in prediction markets was rated by far most difficult. It appears as if high perceived difficulty was crucial for the poor ratings on other categories. In turn, this could form obstacles for the further adoption of prediction markets by practitioners. Researchers and developers should further increase their efforts to make market software solutions more accessible to non-experienced traders.

Participants' perceptions of a group process can be crucial for the acceptance of its results. When agreement among – and satisfaction of – group members is high, decision-makers can share responsibility for their decisions. Thus, the results may entice decision-makers to rely on methods involving personal communication. However, such a strategy can bring along drawbacks.

- There is no evidence that high perceived satisfaction is correlated to good performance.

- Meetings can lead to poor decisions (Armstrong 2006). As shown in Chapter 6, FTF were inferior for questions where the majority of group members had relevant information; a situation that can be expected to be most common when making group decisions in organizations. If one decides to conduct a meeting, one will only invite people who are expected to be able to contribute relevant information.

- Taking into account administrative and time efforts, meetings are expensive and can be difficult to conduct as they require participants' presence.

- Meetings are limited regarding the number of participants. In contrast, Delphi and prediction markets do not require presence of participants and can be conducted asymmetrically with a large number of people.

7.5 Summary

This chapter reported on the evaluation of an ex-post questionnaire from the laboratory experiment described in Section 3.4. In particular, participants' perceptions of the group and the group process on a whole were analyzed. Methods involving personal communication (FTF and NGT) were rated most favorable. Prediction markets were rated least favorable, particularly in terms of difficulty of participation. We attributed this to participants' unfamiliarity with the approach. The results suggest that researchers and developers should further increase their efforts to make market software solutions more accessible and easier to understand for non-experienced participants.

Chapter 8

Advice-taking from Prediction Markets, Meetings, and the Delphi Method

Few decisions are made in isolation. Before making a decision, people often seek advice and aim at gaining more or new information, in particular when dealing with important or risky situations. Furthermore, in seeking advice, decision-makers share accountability for the outcome of decisions (Harvey & Fischer 1997). In the literature, the behavior of seeking advice is referred to as *advice-taking*, *interactive decision making* or *Judge-Advisor Systems* (hereafter, 'JAS'). In the latter, the term judge refers to the person receiving advice from the advisor(s). The judge has to decide how to use the advice and is responsible for making the final decision.

JAS studies typically focus on situations in which (a) the advisors act independently from each other and / or (b) the judge does not interact directly with the advisors. To date, no study is known that analyzed full and simultaneous interaction between judges and a group of advisors. Yet, this type of interactive decision-making is common in real-world meetings. For example, in the process of hiring a new employee, a manager might meet up with his team members to ask for their opinion on different applicants before making the final

decision alone. Similarly, the applicant may consult her family members or friends about which job offer to accept.

As shown in Chapter 7, people's perceptions of prediction markets appeared to be poor, which might result in low confidence in the market results. As a result, decision-makers might hesitate to rely on market results but are likely to revise them. Thus, the question arises how people should use market results and whether they can increase decision accuracy by revising them.

This chapter analyzes data from the laboratory experiment described in Section 3.4. In particular, it compares three JAS variants in which judges and advisors fully and simultaneously interacted with each other. In each variant, judges had to solve the same 10 quantitative judgment tasks. The structure and amount of interaction differed between the variants: group interaction was conducted either as a traditional face-to-face meeting, a Delphi study, or a prediction market. For each variant, it was analyzed whether judges discounted advice and how this affected decision accuracy.

8.1 JAS design

According to the framework proposed in the literature review by Bonaccio and Dalal (2006), a JAS can be described by three categories: input, process, and output. The present study analyzed three variants of a JAS that differed mainly in terms of the process category. This category defines the amount and structure of interaction permitted between advisors and judges. In general, the amount of interaction can vary along a continuum ranging from no interaction at all to full interaction between advisors and judges.

Here, in each of the three JAS variants, the judges fully and simultaneously interacted with a group of advisors in a group process, either in person or computer-mediated. After the group interaction, the judge received the group result as advice for his final individual

decision. Yet, the group process in the three JAS variants differed in the amount and structure of interaction permitted between advisors and judges. In particular, group interaction was conducted either as a traditional face-to-face meeting, a Delphi study, or a prediction market.

8.2 Related Work

In their review of the JAS literature, Bonaccio and Dalal (2006) found that using advice tends to improve decision accuracy. The reason for this goes back to the well-established principle of combining (Armstrong 2001): Yaniv and others (Yaniv 2004a, Yaniv 2004b, Yaniv & Kleinberger 2000) argued that combining diverse opinions reduces the random error associated with each individual judgment and, thus, increases accuracy.

However, findings from the literature show that people do not use advice nearly as much as they should. Instead, they tend to overweight their own opinion, a behavior referred to as advice discounting (Bonaccio & Dalal 2006). Harvey and Fischer (1997) found that judges move their initial estimates only by an amount of 20% to 30% towards the advice.

Advice discounting has been accounted to three causes: *anchoring*, *differential information*, and *egocentric bias*. In the anchoring explanation, a judge's initial estimate may have served as an anchor and, therefore, was inadequately adjusted, according to the advice (Tversky & Kahneman 1974). According to the differential information explanation (Yaniv 2004a, Yaniv 2004b), judges do not have access to the advisors' reasoning and, therefore, have few evidence to justify the advisors' decision. Rather, they rely on their internal justifications and reasons. Finally, Bonaccio and Dalal (2006, p.129) found egocentric advice discounting to be "one of the most robust findings in the JAS literature". Thereafter, judges prefer their own opinion, simply because they believe it to be superior to those of others.

In addition to these three common explanations, a fourth one seems to be plausible: *misconception of combining*. In a series of experiments with MBA students from an elite institution, Larrick and Soll (2006) showed that – even intelligent – people did not understand the power of combining judgments. Instead, many people believed that averaging judgments equals average performance. As a result, people fail in harnessing the benefits of combining when revising judgments.

Yet, despite people's failure in appropriately using advice, there is some good news. In reviewing the literature, Bonaccio and Dalal (2006) concluded that judges' ability to distinguish between good and bad advice increases when the amount of decision-relevant information increases. As a result, they discount poor advice more heavily than good advice.

8.3 Hypotheses

When given advice on a quantitative judgment task, a judge has two possibilities: He can either fully utilize the advice or discount it. Advice utilization refers to the extent to which a judge follows the advice and advice discounting refers to the extent to which the judge does not use the advice. If the goal is to come up with the most accurate estimate, a judge should only discount advice if he thinks he can do better.

The questions addressed in this study were *how* judges use advice from a group interaction and *whether* they can improve a group's decision by discounting advice. Furthermore, it was analyzed whether the results differed depending on the type of group interaction.

Traditional studies on advice-taking mostly analyzed 'prototypical' JAS. In such studies, participants enter the laboratory and are assigned to either the role of a judge or advisor. Furthermore, the advisors usually act independently from each other and do not interact directly. In contrast, the present study focused on advice discounting after the judge simultaneously interacted with a group of advisors. This is a common situation in real-world decision-making: for example, a decision-maker might conduct a meeting and use its

result as input for his final individual decision. Despite these differences, the results were expected to confirm the main findings from the literature: judges should tend to discount advice.

H_1: Judges will discount advice more often than fully utilize it.

Practical experience with prediction markets indicates that people's trust in the results is low. Although we know that a large number of companies are experimenting with prediction markets, published case studies are few. One reason might be a lack of understanding of how prediction markets work. As described in Section 1.2.2, people have problems in translating information into market prices and do not understand how to read them. Furthermore, the analysis in Chapter 7 showed participants' unfavorable perceptions of prediction markets compared to meetings, the Delphi method, and nominal groups. Overall, prediction market participants were less satisfied with the process and participation in markets was perceived as clearly most difficult.

Given the lack of understanding and the unfavorable perceptions of prediction markets, PM judges were assumed to have lower trust in market outcomes than participants in Delphi and meetings. In addition, PM judges did not have access to the advisors' reasoning. Although prediction market participants exchanged information continuously through trading, they could not provide reasoning for their estimates. On the contrary, particularly in the case of FTF (and partly in Delphi through written comments), FTF and Delphi judges had access to the advisors' reasoning. Thus, they could not only gain new or more information and receive feedback on their initial views. They could also get insight into how the group arrived at the decision. If a judge trusts the advice, one would expect him to rely on it. Vice versa, the lower his trust, the higher should be advice discounting. It was expected that:

H_2: PM judges will discount advice more *often* and more *heavily* than FTF and Delphi judges.

Furthermore, since using advice has generally been found to increase decision accuracy, it was expected that:

H_3: Advice discounting by judges will not improve decision accuracy.

Finally, judges' estimates for each group were combined (CJ) and compared to the respective advice. In practice, one could imagine conducting a meeting to discuss a problem. After the meeting, each participant would independently reveal a final individual estimate. The final group result would then be the aggregated outcome of the final individual estimates. In the literature, such approaches have been studied as *talk-estimate* (Gustafson et al. 1973), which is closely related to the nominal group technique. Also, this is related to the Delphi method, as the results from the final round could be interpreted as combined judges' estimates after receiving results from the previous round. As reported by Rowe and Wright (2001), several studies have shown higher accuracy of Delphi results with increasing number of rounds. In general, structured combined estimates have been found to be more accurate than those by individuals (Armstrong 2001). Thus, although individual advice discounting was expected to harm accuracy, it was hypothesized that this effect would be compensated by CJ:

H_4: CJ will improve decision accuracy.

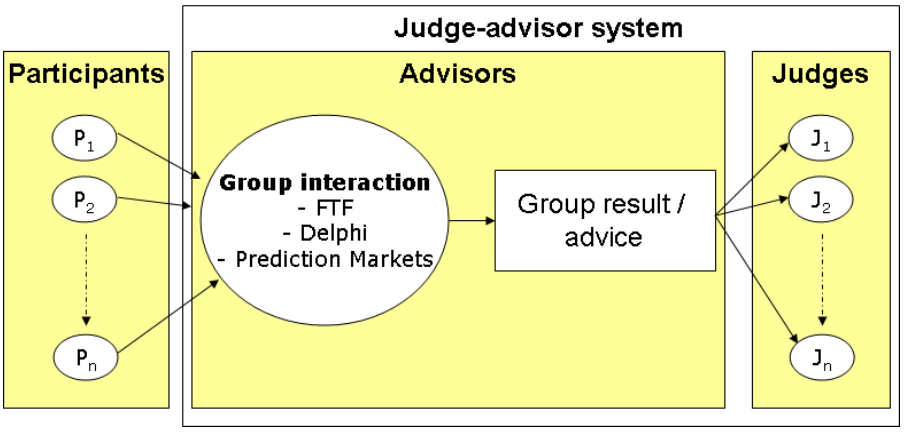
8.4 Study Design

This chapter analyzes part of the data from the laboratory experiment described in Section 3.4. The study design is shown in Figure 9. Participants (n=160) were randomly assigned to

one of 31 groups – 11 FTF, 9 Delphi, and 11 prediction markets.[51] Most groups (30 out of 31) consisted of 4 to 6 subjects; one prediction markets group consisted of 7 subjects.

In a first step, each participant had the role of an advisor by participating in the respective group interaction process. The goal of the group interaction was to come up with group results for 10 quantitative judgment tasks. Thereby, judges were asked to note the achieved group results and to state their confidence in the group results. Then, they were asked to provide their final individual estimates, again along with confidence assessments for each of the 10 tasks. All confidence assessments were revealed on a 7-point Likert scale (1: not at all confident; 7: extremely confident). Examples of the questionnaires are provided in Appendix M.2.

Figure 9: Study Design of the Judge-advisor System



8.5 Results

The data that has been analyzed in this chapter can be found online.[52] Again, as in Chapter 6, the APE was used as a measure of accuracy.[53]

[51] Two Delphi groups had to be discarded as no posterior estimates were obtained. Furthermore, no posterior estimates could be obtained from the NGT groups as the results had to be calculated manually and, thus, were not available at the end of the group process.

- Advice error was measured as the APE of the group result compared to the correct answer.
- Judge error was measured as the APE of the judge's final estimate compared to the correct answer.
- Advice discounting was measured as the APE of judge's final estimate relative to the group result (i.e. advice). A value of zero meant no difference between the judge's estimate and the advice and, therefore, full advice utilization. The higher the value, the higher was the difference between the judge's estimate and the advice. Thus, the higher was advice discounting.
- For measuring the relative accuracy of the advice and the judge's estimate, the error reduction (i.e. the difference between advice error and judge error) was calculated. A positive (negative) value of error reduction means that the judge's estimate improved (harmed) the group result.

8.5.1 Frequency and magnitude of advice discounting

Table 16 shows the overall number of judges' estimates as well as how often judges discounted or fully utilized the group estimate (in the following referred to as advice). Overall, 1,596 judges' estimates were obtained: 520 from the FTF variant, 470 from the Delphi variant, and 606 from the prediction markets variant. As expected (H_1), overall, as well as for each variant, discounting advice was the dominant strategy: in 81% of all cases, judges discounted the advice.[54] The results also corroborated H_2: Prediction market outcomes were discounted far more often (93.2% of the times) than FTF (73.7%) and Delphi

[52] SPSS data file: http://andreas-graefe.org/data/advicetaking.sav

[53] The analysis was replicated by using the error ratio as an index of accuracy. The results were similar. For tests of statistical significance, the data was transformed by taking the square root of all APEs to account for extreme values.

[54] A chi-square test was conducted on the null hypothesis that judges discount advice equally often as they fully utilize it. For each variant, differences were statistically significant ($p<0.01$).

(72.6%). In fact, only 41 times, judges fully utilized the market result whereas 10 of these estimates go back to one participant who did not discount any of the group estimates but fully trusted in the market results. In 565 cases, the PM judges' final estimates differed from the market results.

Table 16: Frequency of advice discounting

JAS variant	Judges' estimates		
	Discounted	Equal to group result	Sum
FTF**	383 (74%)	137 (26%)	520
Delphi**	341 (73%)	129 (27%)	470
PM**	565 (93%)	41 (7%)	606
Overall**	1289 (81%)	307 (19%)	1596

Furthermore, it was expected that PM judges would discount advice more heavily than Delphi and FTF judges. Yet, the FTF (and not the PM) judges discounted advice most heavily. Advice discounting was lowest for Delphi judges.[55]

8.5.2 Advice discounting and accuracy

If a judge discounted advice, there were two possible outcomes: He indeed knew better and his final estimate was closer to the correct answer and, therefore, more accurate than the group result. Alternatively, he did not know better and, consequently, the final estimate was less accurate than the advice.

8.5.2.1 Accuracy of judges' estimates

It was hypothesized that individual advice discounting would not improve accuracy (H_3). Accuracy of the judges' final estimates was analyzed for each question as well as over all 10 questions (per judge).

[55] An ANOVA was conducted to compare the magnitude of advice discounting between the three variants. The results showed significant differences (p=0.02). The ANOVA assumptions were met.

Per question

Each judge estimate was compared to the respective advice. The results, shown in Table 17, met the expectations. For each of the three JAS variants, the majority of judges' estimates did not improve the advice. In fact, judges' final estimates were significantly more often less accurate.[56]

Despite these results, on average, judges could still have improved overall decision accuracy if

- Judges, that were less accurate than the advice, were only slightly less accurate (i.e. small advice discounting in the wrong direction) but
- Judges that were more accurate, were clearly more accurate (large advice discounting in the correct direction)

Table 17: Judges' discounted estimates compared to advice per question

JAS variant	Accuracy of judges compared to advice			Sum
	Higher	Lower	Equal	
FTF*	171 (45%)	212 (55%)	0	383
Delphi*	150 (44%)	191 (56%)	0	341
PM*	257 (46%)	307 (54%)	1[57]	565
Overall*	578 (45%)	710 (55%)	1	1289

To address this, the medians of advice error and judge error were compared per question. As shown in Table 18, on average, advice discounting increased the error for each variant,

[56] A chi-square test was conducted on the null hypothesis that judges' estimates were equally often more than less accurate. For each variant, differences were statistically significant ($p<0.05$). Thus, the null hypothesis was discarded.

[57] In the PM variant, one judge was equally accurate than the advice, even though he discounted the advice. The reason for the identical error is that his final estimate underestimated the correct answer by the same amount that it was overestimated by the advice.

although for FTF and Delphi differences were small. Only in the case of prediction markets, advice discounting significantly harmed decision accuracy.[58]

Table 18: Advice error vs. judge error per question

JAS variant	N	Median of		Error reduction
		Advice error	Judge error	
FTF	383	0.32	0.33	-0.01
Delphi	341	0.21	0.25	-0.04
PM*	565	0.26	0.30	-0.04
Overall	1289	1.58	1.66	-0.08

Per judge

For each judge, median judge error and advice error were aggregated over all 10 questions and compared in terms of accuracy. The results are shown in Table 19. For FTF and Delphi, more judges were more accurate than less accurate, although differences were small. In the case of prediction markets, advice discounting harmed decision accuracy: only 22 judges were more accurate than their group, whereas 38 were less accurate. These differences were significant: over all 10 questions, the majority of PM judges were less accurate than the advice.[59]

Table 19: Judges vs. advice over all 10 questions

JAS variant	Accuracy of judges compared to advice			Sum
	Higher	Lower	Equal	
FTF	26 (50%)	24 (46%)	2 (4%)	52
Delphi	25 (53%)	18 (38%)	4 (9%)	47
PM*	22 (36%)	38 (62%)	1 (2%)	61
Overall	73 (46%)	80 (50%)	7 (4%)	160

For each judge, advice error and judge error were compared over the 10 questions. The results are shown in Table 20. For FTF and Delphi, differences were small. By comparison,

[58] For each variant, a paired t-test was conducted to compare the means of advice error and judge error.

[59] Chi-square tests were conducted on the null hypothesis that judges would be equally often more than less accurate.

in the case of prediction markets, judges were significantly less accurate than their group: when discounting advice, PM judges harmed the results more than FTF and Delphi judges.[60]

Table 20: Advice error vs. judge error over all 10 questions

JAS variant	N	Median of		Error reduction
		Median advice error	Median judge error	
FTF	52	0.32	0.30	0.02
Delphi	47	0.19	0.21	-0.02
PM*	61	0.24	0.29	-0.05
Overall*	160	0.26	0.26	0

8.5.2.2 Accuracy of combined judgments (CJ)

It was hypothesized that combining the judges' estimates (CJ) would improve decision accuracy (H_4). Again, accuracy was analyzed per question as well as over all 10 questions (per group).

Table 21: CJ vs. advice per question

JAS variant	Accuracy of CJs compared to advice			Sum
	Higher	Lower	Equal	
FTF**	41 (37%)	39 (36%)	30 (27%)	110
Delphi**	30 (33%)	22 (25%)	38 (42%)	90
PM	49 (44%)	45 (41%)	16 (15%)	110
Overall**	120 (39%)	106 (34%)	84 (27%)	310

Per question

For each question, the judges' estimates of the same group were combined (CJ) by using the median. Then, the CJ errors were calculated and compared to the respective advice. The results are shown in Table 21. For prediction markets, differences were small. For FTF and Delphi, the results showed significant differences.[61] However, the results did not meet the

[60] Paired t-tests were conducted to compare advice error and judge error per variant.

[61] Chi-square tests were conducted on the null hypothesis that – compared to the advice – the CJs were equally often more than less or equally accurate.

expectations: the majority of CJs did not improve the advice. In addition, for each variant, differences in the magnitude of advice error and CJ error were small.[62]

Per group

For each group, the median advice error and median CJ error were compared over all 10 questions. The results are shown in Table 22 and Table 23. Only for prediction markets, less accurate CJs outnumbered more accurate ones. For FTF and Delphi, a majority of CJs was more accurate than the advice over all 10 questions.[63]

Table 22: CJ vs. advice over all 10 questions

JAS variant	Accuracy of CJs compared to advice			Sum
	Higher	Lower	Equal	
FTF	8	2	1	11
Delphi	7	2	-	9
PM	4	6	1	11
Overall	19	10	2	31

Accordingly, as shown in Table 23, CJs over all 10 questions reduced error in the case of FTF and Delphi, although differences were significant only for Delphi.[64] In the case of prediction markets, the CJs did not improve decision accuracy.

Table 23: Advice vs. CJ error over all 10 questions

JAS variant	N	Median of		Error reduction
		Median advice error	Median CJ error	
FTF	11	0.32	0.29	0.03
Delphi*	9	0.19	0.18	0.01
PM	11	0.24	0.24	0
Overall	31	0.28	0.24	0.04

[62] A paired t-test showed no statistically significant differences for advice error and CJ error for any of the variants.

[63]Chi-square tests were conducted on the null hypothesis that CJs were equally often more accurate than less or equally accurate (compared to the advice). Due to the small number of observations, differences were small for each variant.

[64] Paired t-tests were conducted to compare advice error and CJ error for each variant.

8.5.3 Correlations

Cases in which judges did not discount advice were excluded for the calculation of the correlation coefficients.

8.5.3.1 Advice discounting

In the following, the relationship of advice discounting to advice error and to error reduction was analyzed. The results are shown in Table 24. Not surprisingly, negative correlations were obtained between advice discounting and error reduction for each variant: the more judges discounted the group result, the more inaccurate were their final individual estimates. This effect was strongest for Delphi and PM. Furthermore, in the case of FTF, advice discounting was positively correlated to advice error. This conforms to findings from the literature that judges discount poor advice more heavily than good advice. However, no such effect could be observed for Delphi and PM. It seems as if FTF judges could differentiate between good and bad advice whereas Delphi and PM judges could not.

Table 24: Correlations of advice discounting vs. advice error and error reduction

JAS variant	Advice discounting correlated to	
	Advice error	Error reduction
FTF	0.17**	-0.47**
Delphi	0.03	-0.62**
PM	-0.02	-0.63**
Overall	0.15	-0.54**

The results can be explained by the differential information explanation for advice discounting. In FTF, judges fully interacted with advisors and, thus, had access to the reasoning of fellow group members. This might have allowed them to determine the reliability of the group result. As a result, they discounted bad advice more heavily. In contrast, Delphi and PM judges might have had problems in determining the quality of the advice. In Delphi, participants had access to the reasoning of others only if fellow group members provided written comments in round one. In the case of prediction markets, participants exchanged information only through trading but were unable to provide

reasoning for their estimates. They could only buy and sell contracts, without indicating why they did so.

8.5.3.2 Confidence

Table 25 reports on judges' confidence in the advice and in their final individual estimates.[65] Both times, Delphi judges were most confident, followed by PM judges. FTF judges were least confident. Furthermore, for each variant, judges were more confident in their final individual estimate than in the advice.[66]

Table 25: Judges' confidence in advice and final estimates

JAS variant	N	Mean confidence in		Confidence increase
		Advice	Final estimate	
FTF**	375	3.7	3.9	0.2
Delphi**	320	4.2	4.6	0.4
PM*	563	3.9	4.1	0.2
Overall**	1258	3.9	4.2	0.3

In the following, *absolute* confidence refers to the judges' confidence in their final individual estimates. The difference between confidence in the individual estimate and the advice is referred to as *confidence increase*. Confidence increase is positive (negative), when a judge was more (less) confident in his final individual decision than in the advice. It was analyzed how judges' absolute confidence and confidence increase were related to error reduction, advice discounting, and judge error.

The results are shown in Table 26. In general, correlations were either not existent or small:

[65] Again, cases in which judges did not discount advice were excluded. Some participants did not reveal confidence scores at all or for some questions. Thus, the number of observations for confidence scores was not identical with the number of estimates.

[66] For each variant, paired t-test were conducted to compare the confidence in advice and in the final estimate.

- There were no correlations between confidence and judge error: more confident judges were not more accurate.
- Not surprisingly, for FTF and prediction markets, small correlations could be identified between confidence increase and advice discounting: judges who were more confident in their individual estimate than in the group estimate discounted advice more heavily.
- There was also a small negative correlation between confidence increase and error reduction for prediction markets: although PM judges were more confident in their final estimates, their revised estimates harmed accuracy.
- In Delphi, absolute confidence was positively correlated to error reduction. Delphi judges who were more confident in their final estimates were also more accurate.
- In the case of FTF, absolute confidence was positively correlated to advice discounting. FTF judges who were more confident in their final estimates discounted advice heavier.

Table 26: Correlations of advice discounting vs. advice quality and error reduction

JAS variant	Absolute confidence correlated to			Confidence increase correlated to		
	Error reduction	Advice discounting	Judge error	Error reduction	Advice discounting	Judge error
FTF	-0.07	0.21**	0.04	-0.06	0.17**	0.07
Delphi	0.14**	-0.09	-0.04	0.04	0.1	0.02
PM	0.02	-0.01	-0.07	-0.10**	0.11**	0.03
Overall	0.02	0	-0.04	-0.05	0.11**	0.04

8.6 Discussion

This study examined three variants in which judges simultaneously interacted with a group of advisors. The results conformed to findings from the JAS literature, whereupon advice-taking has been found to increase decision accuracy (Bonaccio & Dalal 2006). In each variant, advice discounting tended to harm accuracy. Thus, the results do not support the way decisions are often made in the 'real world' where decision-makers often meet up with others to ask for additional views before making the final decision alone.

Furthermore, there were hardly any differences between structured and computer-mediated (Delphi and PM) and unstructured, in-person approaches (FTF). However, the results somewhat differed for the prediction markets variant. PM judges discounted advice far more often than Delphi and FTF judges. In addition, advice discounting in prediction markets harmed decision accuracy more than in the cases of FTF and Delphi. Per question as well as per judge, PM judges' estimates did worse than the advice whereas differences in the case of Delphi and FTF were small.

A reason might be that people had unfavorable perceptions of prediction markets and, therefore, trust in prediction market outcomes was low. In practice, this could lead to a general resistance against the implementation of prediction markets. At the least, decision-makers are likely to discount market results instead of fully utilizing them – a strategy that would harm decision accuracy.

Yet, low confidence in market outcomes cannot have been responsible alone. The analysis of judges' confidence revealed that PM judges were more confident than FTF judges. This supports the differential information explanation for advice discounting. In FTF, judges can derive from discussion what information is embodied in the group result. As shown by the correlation coefficients, direct interaction might have helped FTF judges to obtain more decision-relevant information and, thus, to differentiate good and bad advice. As a result, FTF judges might have discounted bad advice particularly strong (advice discounting was heaviest in FTF). On the contrary, the results corroborate the assumption that people have problems in understanding how prediction markets work. In addition, they cannot derive the reasons why others buy or sell a particular contract. This makes it difficult to assess the quality of results and, thus, to appropriately discount (bad) advice.

Furthermore, even combining judges' estimates did not outperform the advice. For none of the variants, the majority of CJs were more accurate than the advice per question. Yet, for Delphi, the CJs – which are essentially an additional Delphi round – appeared to

outperform the advice per group. This conforms to findings of Rowe and Wright (2001), who reported on several studies that have shown higher accuracy of Delphi results with increasing number of rounds

8.7 Summary

This chapter analyzed data from the laboratory experiment described in Section 3.4. Thereby, meeting, Delphi, and prediction market groups were examined as JAS systems. After being presented with the results from the group interaction, participants (judges) were asked to provide their final individual estimates on each of the ten questions.

The results confirmed findings from the JAS literature: In general, judges tended to discount the group result (advice). Judges, that were more confident in their individual estimate than in the group result, discounted advice more heavily. In FTF, bad advice was discounted more heavily than good advice, although this was not the case in Delphi and prediction markets. Finally, advice discounting generally harmed decision accuracy.

The results suggest that the way decisions are often made in the 'real world', when people consult others to seek for advice before making the final decision alone, does not lead to the maximum achievable levels of accuracy. When interacting with a group, decision-makers should refrain from discounting the group result.

In addition, differences between the three JAS variants were identified. In the case of Delphi and FTF, differences in the quality of judges' estimates and the advice were small. On the contrary, PM judges harmed decision accuracy when discounting advice. Nonetheless, PM judges discounted advice clearly more often than FTF or Delphi judges. These results appear to be consistent with practical experience, indicating that people have problems in understanding how prediction markets work. They also conform to the findings from Chapter 7, showing that participants had comparably unfavorable perceptions of prediction

markets. Again, this might represent a barrier for the implementation of prediction markets as decision-making tools, which is unfortunate as people fail in improving market accuracy by discounting the results.

Chapter 9

Summary

This work was motivated by the need to analyze some of the remaining barriers that hinder the implementation of prediction markets in practice. At the time of writing, no organization is known that uses prediction markets to an extent that goes beyond experimental status. Prediction markets have yet to become an established forecasting method. The reasons for this are manifold.

Although the studies available to date demonstrate high forecasting performance of prediction markets in various fields, their number is still limited and they are often small scale. Since the emergence of the field, no meta-analysis has been published to analyze prediction markets' accuracy. In particular, little is known about the relative performance of prediction markets compared to established structured judgmental approaches. As long as there is no evidence that prediction markets are superior to alternative mechanisms, organizations have no need to depart from their status quo. Furthermore, the majority of studies focused on predicting events in the near future. There is a lack of studies that analyze the applicability of prediction markets for forecasting events in the distant future or similar problems like long-term trends.

In addition, practical experience indicates that prediction markets face cognitive and organizational barriers. People seem to have problems in understanding how prediction markets work. They do not understand how to reveal information through trading or how to interpret market prices as forecasts. Besides that, prediction markets change the way decisions are traditionally made in organizations: in using a prediction market to obtain a forecast, one also makes the forecast publicly available. This can meet with a refusal by decision-makers as it makes wrong decisions transparent and reveals accountability. Furthermore, prediction markets do not necessarily aim at limiting participation to potential experts but are usually open for everyone to participate. Despite evidence that disputes the value of experts in forecasting and supports the potential of aggregating dispersed knowledge, people fear that involving a large number of amateurs may harm forecasting accuracy.

In analyzing data from two empirical studies, this work addressed some of the questions that arise from these implementation barriers for prediction markets.

9.1 Contributions of this Work

In providing the first empirical comparison of traditional meetings, nominal groups, the Delphi method and prediction markets on a quantitative judgment task, the present work made contributions to the forecasting and decision-making literature that go beyond the research field of prediction markets. The design of this laboratory experiment was presented in Section 3.4. The findings can be briefly summarized as follows.

- Chapter 6 compared the outcomes of the four group methods in terms of accuracy. As expected, overall, the three structured approaches (prediction markets, nominal groups, and Delphi) outperformed meetings, although differences were small. Furthermore, meetings did not generally perform worst but the methods differed according to the difficulty of questions. The three structured approaches

outperformed meetings on easy-to-estimate questions for which one could expect the majority of group members to possess relevant information. By comparison, meetings tended to be superior for hard-to-estimate questions for which only few group members might have had relevant information. The reason might have been that, for such problems, direct interaction in meetings enabled well-informed individuals to convince fellow group members of their beliefs. In Delphi, and particularly prediction markets, the amount of interaction and the ability to provide reasoning for one's judgment is limited. Thus, it becomes hard – if not impossible – for well-informed participants to convince the group if the majority of members are poorly informed.

- After being presented with the group outcomes, participants of meetings, Delphi, and prediction markets were asked to provide their final individual estimates. Such studies are known as judge-advisor-systems. The results, reported in Chapter 8, conformed to findings from the literature. In general, individuals tended to discount the group result. Individuals, that were more confident in their individual estimate than in the group result, discounted the group result more heavily. In meetings, bad advice was discounted more heavily than good advice, although this was not the case in Delphi and prediction markets. The reason might have been that meeting participants (in contrast to Delphi and prediction market participants) had better access to the reasoning of their fellow group members, which helped them to judge the quality of the group result. Finally, advice discounting generally harmed decision accuracy. The results suggest that the way decisions are often made in the 'real world', when people meet up with others to seek for advice before making the final decision alone, does not lead to the maximum achievable levels of accuracy. When interacting with a group, decision-makers should refrain from discounting the group result.

- Finally, participants of each method were asked how they perceived their group as well as the group process as a whole. The results were presented in Chapter 7. While methods involving personal interaction (meetings and nominal groups) were rated favorable, prediction markets were rated most unfavorable, particularly in terms of difficulty. This conformed to findings from the literature, showing that people enjoy human interaction and the sense of working together. By comparison, the unfavorable rating of prediction markets might have been caused by the novelty of the approach and people's lack of understanding of how prediction markets work.

However, the main goal of this work was to address the five research questions raised in Chapter 1, which arose from remaining barriers that hinder the implementation of prediction markets in practice. Given the results of this work, these can be answered briefly as follows.

Research question 1:

How do prediction markets perform for long-term forecasting?

In reporting on data from the TechForX field experiment, the first empirical study that compared prediction markets to the Delphi method, this question was analyzed in Chapter 4. The results of two prediction markets – one comprised of experts and one of students – were compared to the results of a large-scale Delphi study. The task was a classical judgment task: participants had to come up with assessments of long-term trends, which could not be judged upon their correctness. Both markets generated outcomes that strongly correlated with the Delphi results. Thereby, the level of correlation with established methods was consistent with findings from earlier studies in the field of market research that analyzed prediction markets for the same type of tasks. It was concluded that prediction markets can be valuable for solving judgment tasks whose outcome cannot clearly be judged.

Research question 2:

Do experts have value in prediction markets?

This question was analyzed in Chapter 5 by comparing results and trading behavior in the TechForX expert and student markets. The results of both markets were similar compared to the Delphi results. This conformed to findings from the literature, showing that expertise has little value in forecasting change. However, an analysis of trading behavior revealed differences between experts and students. While experts' confidence, measured as the trading volume, varied along different topics, students' confidence did not differ. Furthermore, experts were more confident for hard- and moderate-to-assess forecasting environments, whereas this was not the case for easy-to-assess forecasting environments. Also, experts were not generally more confident than students. This led to the conclusion that experts revealed their beliefs well-considered, whereas students appeared to 'just trade'. Thus, the results suggested that experts do have value in prediction markets. Furthermore, in analyzing trading behavior, one can reveal information about people's subconscious confidence. This can increase validity and reliability of results. That way, prediction markets can improve on traditional approaches.

Research question 3:

How do prediction markets perform compared to traditional approaches?

The results from the TechForX field experiment (Chapter 4) suggested that prediction markets perform equally to Delphi for long-term forecasting problems without clear outcomes. In Chapter 6, prediction markets were compared to meetings, nominal groups and Delphi on a quantitative judgment task, consisting of ten factual questions with demonstrably correct solutions. Over all ten questions, prediction markets performed equally to the three alternatives and there was generally little difference between the three

structured approaches (prediction markets, nominal groups, and Delphi). However, differences could be identified compared to meetings. While prediction markets outperformed meetings on easy-to-estimate questions, meetings tended to be superior for hard-to-estimate questions. It was concluded that the limited amount of interaction permitted between prediction market participants was responsible for their poor performance on hard-to-estimate questions. If people cannot provide reasoning for their judgment, well-informed participants might fail in convincing the group of their information. By comparison, prediction markets appeared to be particularly valuable in situations where many group members have valid information on the question.

Research question 4:

How do people perceive participation in prediction markets?

This question was addressed in different parts of this work. The comparison of participants' perceptions of the different methods, reported in Chapter 7, showed that prediction markets were rated most unfavorable on the majority of categories, in particular in terms of difficulty, freedom to participate, and satisfaction. These poor ratings conformed to practical experience indicating people's cognitive challenges when using prediction markets. The findings were corroborated by the analysis of how individuals made use of the group results when asked to provide their final individual estimates (Chapter 8). Compared to Delphi and meeting groups, prediction markets participants discounted the group results clearly more often, which, again, can be interpreted as an indication for low confidence in the market results. In sum, the results suggested that prediction markets face a barrier of low acceptability, which might be one explanation for why the method has not been more widely adopted in practice yet.

Research question 5

How do people use market results? How should they use them?

The results of the advice-taking study, reported in Chapter 8, showed a clear tendency of participants to revise the market results. However, when revising the market results, participants did not improve forecasting accuracy; they harmed it. This was the case for individual as well as for combined revised estimates. In addition, it was found that prediction market participants were unable to differentiate good and bad advice. Also, participants who were more confident in their final individual estimate than in the market results were not more accurate. The results suggested that people should trust in prediction market outcomes and should refrain from revising them when making decisions.

9.2 Potentials for Future Research

It is one of the main conclusions from this work that prediction markets face manifest barriers, which keep preventing their implementation as an established forecasting method in organizations. In particular, people appeared to have poor perceptions of participation in prediction markets and, thus, have little trust in the market results.

This is unfortunate, as the available studies to date show that prediction markets perform at least as well as alternative methods – a finding that was corroborated by the two empirical studies in this work. In addition, as described in Section 2.4, prediction markets have potentials to improve on existing approaches. They can involve a large number of people, motivate participation, and aggregate dispersed information fast and continuously. Furthermore, as shown in this work, analyses of trading behavior can reveal additional information (like subconscious confidence levels), which can be valuable for decision-makers.

Further research is necessary to advance the field of prediction markets and to make the method more appealing for organizations. Two streams of research appear to be particularly promising: market engineering and further empirical analyses.

9.2.1 Market engineering

Since the emergence of prediction markets, market engineering has been an important stream of research. As summarized in Section 1.1, thus far, much attention has been focused on finding ways to deal with well-known problems like market inefficiencies or trader biases. Thereby, markets have proven to perform well despite inefficiencies and biased traders. Also, the research available to date could not identify clear differences in accuracy between real-money and play-money. The same applies for market size. In sum, regardless of the specific market design (or the field of application), the studies available to date have shown good forecasting accuracy of prediction markets. That said, without question, if one would be able to identify market inefficiencies, traders' biases, or people's risk attitudes – and could find ways to deal with these problems – one might be able to further enhance the performance of prediction markets.

However, the results from this work revealed potentials for a shift in the focus of market engineering. To deal with the actual barriers that hinder the implementation of prediction markets within organizations, markets developers should look for ways to make their systems more appealing for users. In particular, markets should be more accessible and easier to use for non-experienced traders. Even though the market software used in this work's laboratory experiment was chosen because of its simple user interface, participants experienced prediction markets as clearly more difficult than the other three approaches. In general, the extent of how the user interface affects participants' perceptions of the method has not received much attention in the field of prediction markets yet. It appears promising to conduct experiments with different user interfaces to identify design features that are preferred by participants.

Furthermore, it was shown that the performance of prediction markets decreased for hard-to-estimate problems, for which one could expect only a minority of participants to be well-informed. It was concluded that this is due to limitations in how participants were able to interact with each other: participants could reveal their information only through the process of trading but were unable to provide reasons for their estimates. Thus, it was impossible for well-informed participants to convince the group to adopt their judgment. Future research should analyze situations in which participants are able to communicate with each other, for example, by providing additional means of communication like forums or chat rooms. Although such interaction is common at most commercial prediction market platforms (like intrade.com or hubdub.com), to date there is no research that analyzed its effect on accuracy. While exchanging reasoning might allow for more accurate predictions, it could also harm prediction accuracy as it provides new opportunities for manipulators. Analyzing situations with varying type and amount of interaction between participants seems to be a particularly interesting question for further research.

9.2.2 Empirical analyses

The studies available to date show high accuracy of prediction markets for various fields of application. However, their number is still limited and most are small scale. Furthermore, some studies report only on absolute accuracy (like the correct prediction of a project's finishing date) or compare the results to benchmarks like polls or betting odds. In order to convince decision-makers of the potentials of prediction markets, there is a need for studies that analyze the markets' relative accuracy compared to established forecasting methods used in organizations. Since the emergence of the field, no meta-analysis has been conducted that analyzed prediction markets accuracy for various types of problems and fields of application. While the present work contributed to fill this need, further studies are necessary.

Thereby, it seems valuable to analyze the performance of prediction markets for different types of problems, particularly compared to meetings. Although common practice in organizations, meetings are expensive, time-consuming, and subject to biases that harm forecasting or, more general, decision accuracy. These biases become even more evident in meetings that involve people from different hierarchy levels or address more complicated problems. For example, the quantitative judgment tasks used in this work's laboratory experiment required people to only aggregate information or 'facts'. Group decision-making becomes more challenging for competitive tasks like negotiation as they also involve people's values, attitudes, emotions, expectations, commitments, etc. Future studies should conduct experiments to analyze prediction markets for such tasks, involving people from different hierarchy levels.

9.2.3 Analyses of trading behavior

The results from Chapter 5 revealed potentials in analyzing people's trading behavior. For example, the order volume can be interpreted as a measure for participant's confidence. While such analyses are common in the financial literature, they have not received attention in the field of prediction markets yet. However, such information can be valuable to decision-makers who have to make a decision based on market results. It can be used to alert for uncertainty and can add validity and reliability to the results.

Several research questions appear to be worth investigating. For example, do more confident participants indeed trade higher volumes? If so, are more confident participants also more accurate? Finally, how does providing decision-makers with information about participants' confidence affect the final decision made based on the market results?

9.3 Final Remarks

Prediction markets are a structured judgmental forecasting approach and have been proven to perform well for different types of forecasting problems. However, the approach faces

remaining barriers that hinder its implementation. To overcome these barriers and leverage the further implementation of prediction markets, research in the field remains important. This involves studies in the field of market engineering as well as further empirical work to analyze the relative performance of prediction markets.

Thereby, researchers should refrain from considering this a horse race between prediction markets and traditional means of forecasting. In particular, one should not aim at replacing traditional forecasting methods by prediction markets. Rather, prediction markets should be regarded as a supplemental approach and future research should focus on identifying situations in which they may enhance the field of forecasting. For example, as shown in this work, prediction markets appear to be particularly valuable in situations where new information becomes continuously available and multiple participants have valid insight into the issue in question. Thus, it seems highly promising to combine prediction markets with existing – both quantitative and qualitative – forecasting methods. Such a market platform would allow participants to draw on additional sources of information and, due to the benefits of combining forecasts (Armstrong 2001), make more informed trades. Developing such a platform and identifying ways to combine prediction markets and traditional methods for a particular forecasting problem is up to future research.

Methodological Appendix

M.1 Appendix to the TechForX Field Experiment

M.1.1 Short instructions for participants

These are the short instructions provided to participants in the expert group. Participants in the student group received the same material, translated to German.

Short Instructions

5 theses about expected future developments in the "creative content" sector have been mapped as virtual stocks and will be traded on this prediction market. The theses have been developed by the EPIS Project, where they will be assessed in a Delphi study. At the same time, a subset of the Delphi theses can be traded on the open prediction market TechForX that is accessible for everyone via internet.

In contrast to the open market TechForX, **this experiment is restricted to a number of 25 participants**. It is the goal of this research project to analyze prediction markets as a method for long-term forecasting.

What is your task?
Your task is to assess the probabilities of the theses to come true by trading the respective virtual stocks. The price of a virtual stock always reflects the aggregated probability estimated by all participants for the particular thesis to come true. In addition, it should be your goal to maximize

your depot value in order to climb in the user ranking and to increase your chances to receive a prize.

How much money / How many stocks do you have?
You receive 10.000 TechForX Euros for every market (i.e. for every thesis) and 100 shares of every stock that is traded in the respective market, i.e. you can immediately start with buying and selling stocks.

How does it work?
Example: Assume the situation described in Table 1, i.e. you are asked to judge how likely the thesis ***Online self-publication of books is the predominant way of commercial distribution, even for established authors*** will come true for four different time horizons.

In this example, your estimation that this thesis will *never* (50%) or *not until 2020* (20%) come true exactly matches with the current market prices of the related stocks ***Never*** (50) and ***Realized After 2020*** (20). Thus, if you have the strategy to trade based on your expectations, then you have no reason to trade this stock.

At the same time, you assess the claim to come true *before 2011* as less likely (10%) than its current market price indicates (20). Therefore, you would sell stocks of ***Realized before 2011*** until a price of 10. Analogically, you would buy stocks of ***Realized in 2011-2020*** until a price of 20.

Table 1	Thesis: Online self-publication of books is the predominant way of commercial distribution, even for established authors.	
Stock	**Current market price**	**Your individual estimation**
Never	50	50 %
Realized before 2011	20	10 %
Realized in 2011-2020	10	20 %
Realized After 2020	20	20 %

Important! At the start of the market, i.e. at your first login, the order books are probably still empty or there are no market prices yet. None the less, you can easily reveal your opinion for every thesis by placing two respective orders.
Assume you estimate the probability that the above thesis will never come true between 30 and 40 per cent. Then you would buy the stock ***Never*** until a price of 30, i.e. you place **a buy order at a limit price of 30**. Additionally, you would sell the stock again from a price of 40, i.e. you simultaneously place **a sell order at a limit price of 40**.

Are there different trading strategies?
Yes, there are different strategies to increase your depot value. For example, as described above, you can trade your individual expectations but you have to bear in mind that the final depot value

at the end of the experiment will be determined by the results of the EPIS Delphi study. Furthermore, you may try to get a sentiment for the market and gain from fluctuations of the market price through arbitrage or portfolio trading.

How often are you supposed to trade?
For the experiment to become a success we would like to ask you to **log in at least twice** during the runtime of two weeks and to thereby **trade every stock at least once**. However, in order to maximize your depot value and to climb in the user ranking we recommend to be active in the market more often.

How can you win?
At the end of the experiment an overall user ranking will be calculated from the sum of the depot values of all 5 markets. The particular depot value for each market is thereby calculated from the sum of your money supply and the value of the stocks you own whereas **at the end of the experiment the value of the stocks does not correspond to the last traded market price but is determined by the results of the Delphi study that is conducted in the EPIS Project.** Hence, the user rankings for each market at runtime do not necessarily represent the respective ranking at the end of the experiment. This is illustrated in Table 2.

Table 2 These: Online self-publication of books is the predominant way of commercial distribution, even for established authors.

Aktie	**Result of the EPIS Delphi**	**Number of your stocks**	**Value of your stocks**
Never	30 %	100	30 x 100 = **3.000**
Realized before 2011	30 %	150	**4.500**
Realized in 2011-2020	10 %	200	**8.000**
Realized After 2020	30 %	130	**3.900**
Money supply		**1.500**	
Depot value		**20.900**	

After the experiment the three best-performing participants, i.e. the participants with the highest overall depot values, will be rewarded according to the following pay-off-rule:

1. **Rank (highest overall depot value): 200 €**
2. **Rank (2nd highest overall depot value): 100 €**
3. **Rank (3rd highest overall depot value): 50 €**

Since it will take some time to analyze the results of the parallel running Delphi study, the winners will be notified presumably by the end of July. The results will also be announced on the website and by email to all participants.

Have fun and good luck!

M.1.2 Full tutorial

Tutorial

This short tutorial presents the most important features of ***TechForX*** and is designed to facilitate trading!

Registration

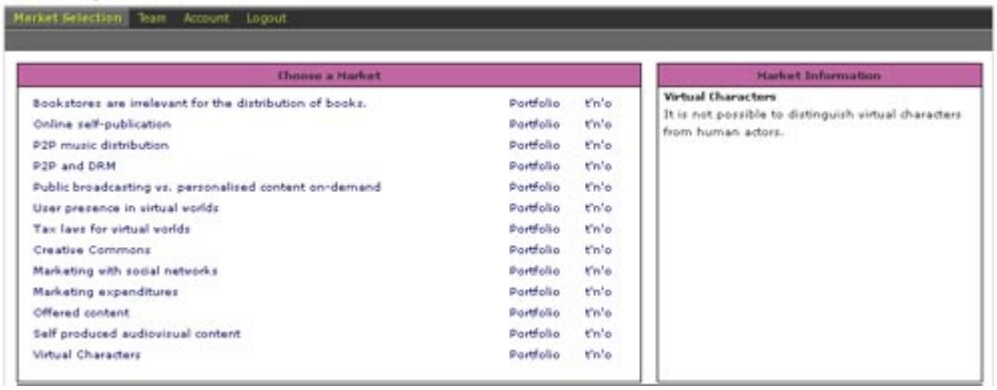
In order to actively participate in ***TechForX***, a *registration* is needed. **Participation is free of charge.**

A personal account will be created for you as soon as you have submitted the complete *registration form*.

After registering, you will receive an e-mail containing a link and a unique *activation code*. This code will activate your chosen username along with the password. Follow the link and activate the access.

Market selection

After the login (or registration, resp.) you can select the desired market you want to trade in.



The market mechanism: CDA

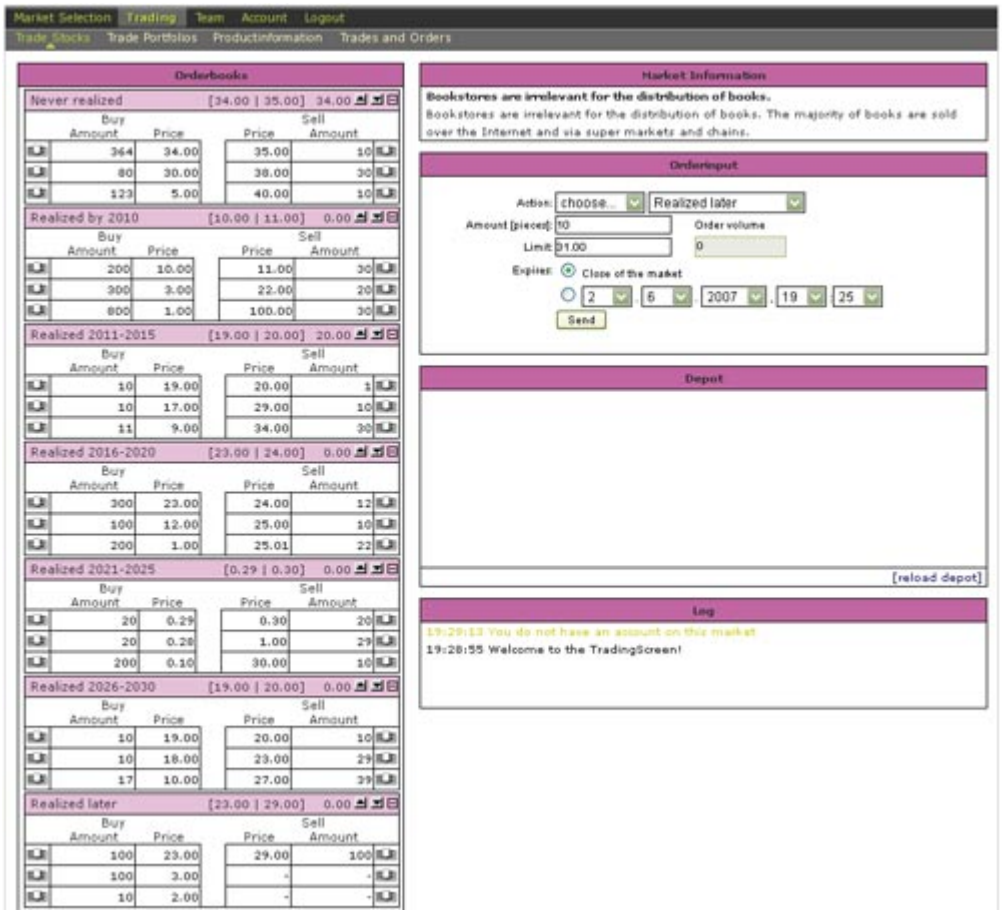
With the **Continuous Double Auction** (CDA), existing orders are executed instantly, as soon as a matching counteroffer is available. A so called price/time priority holds. This means that orders are ordered according to their prices in first priority and according to the time they were received by the system in second priority, independent of their origin or size. If there is no matching counteroffer, the order is placed in the order book. There, the open buy and sell orders are listed according to the price. As soon as the system finds a matching order on the opposite side, the order will be executed instantly and as far as possible.

Endowment

If you register for a market, you will receive an endowment of 100,000 TechForX-Euros.

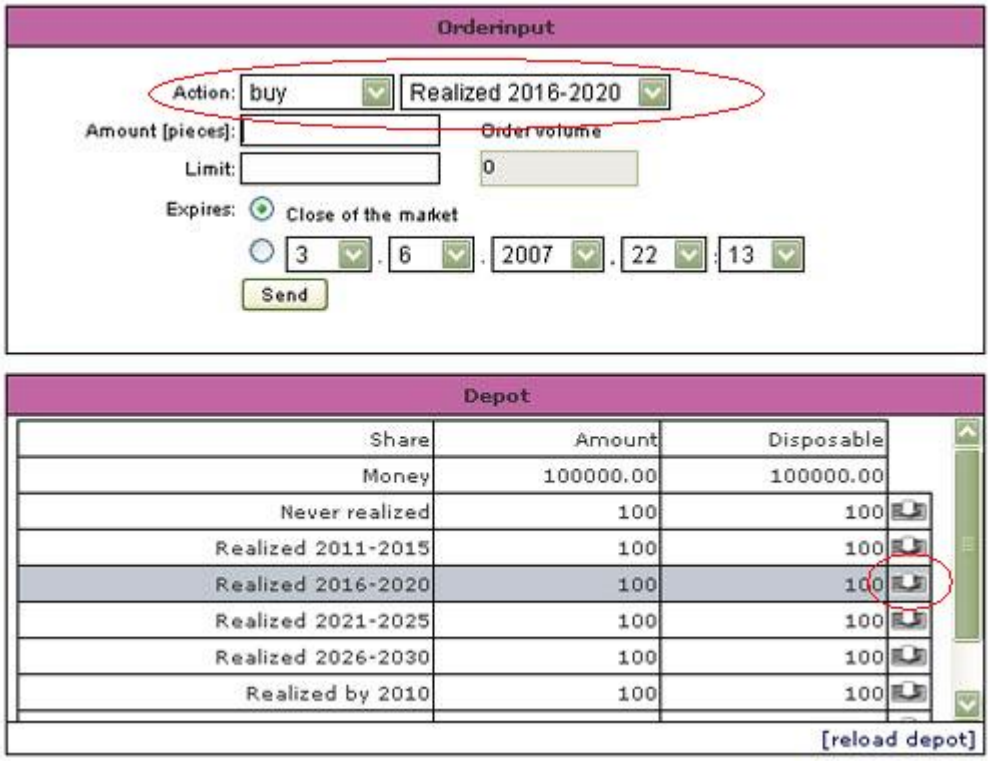
Trading

You can trade by selecting "Trade Stocks" in the menu bar. Besides the *order books* with three current orders of other traders, you can find your current *depot value* and the *mask for the order input (to buy or to sell shares)*. Possible occurring errors are being displayed in the logging area during the order request. (*Messages*).

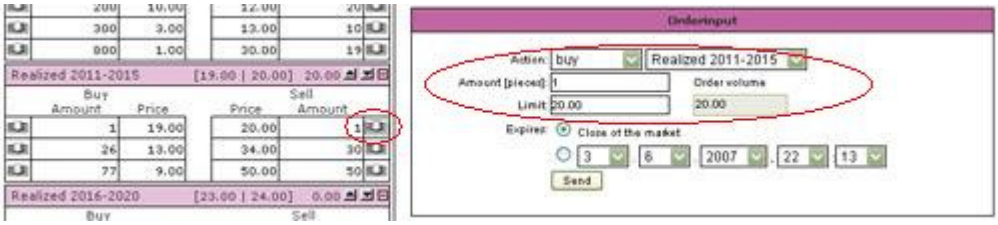


In order to *buy or sell* shares, please select a share and enter the desired quantity and limit price. Additionally, you can specify the period of validity for your order. The order will be executed at that price or below, as soon as there is a matching counterpart on the other side of the market. Your order can only be executed if you have enough shares to sell or, resp., enough money to buy the shares.

Another possibility to trade is the arrow-button. The chosen share will be taken over to the input mask directly.



Example:



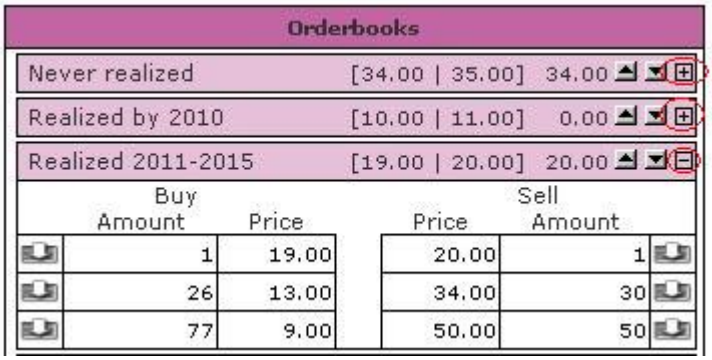
In this example, you directly react to the sell offers of other market participants. In this case, you would buy 1 shares of "Realized 2011-2015" at the price of 20.00.

The order is not valid before you press the button to submit the order. You can change the order, e.g. resetting quantity or limit, before pressing the button. In the same fashion, you can react to sell offers.

Individual arrangement of the order books

You can change the *arrangement of the order books* by clicking the arrows. For instance, the shares you are trading frequently can be arranged at the top.

Additionally, it is possible to fold the order books to gain a better view.

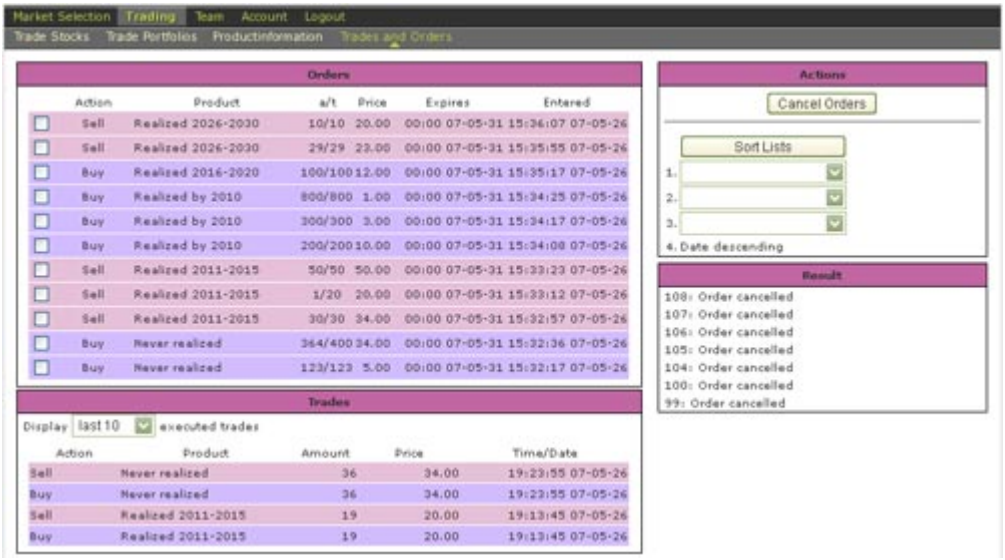


Trades and open orders

In *Trades and Orders* you will find executed as well as open orders.

In row a/t all trades (a) are displayed as well as how many of those have not been executed yet. It is possible to delete the open orders any time.

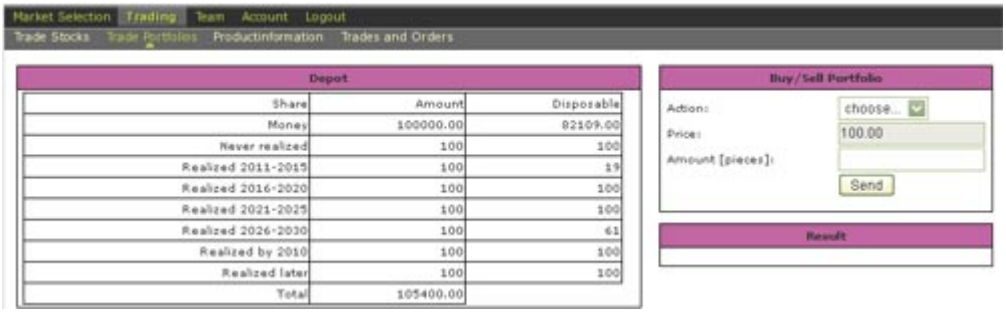
The *Trades* subscreen shows the trades which have been executed.



Portfolio trading

Instead of trading shares with other traders, it is possible to buy and so called unit portfolios under *Trade Portfolios*. A portfolio contains exactly one share of those that can be traded on the current market. The price of a portfolio is set to 100 TechForX-Euros.

You can buy and sell the portfolios in the dialog. A sufficient amount of money/shares is need to buy/sell unit portfolios.



Change password/account information

By clicking Account in the menu bar, you can change your password as well as change personal information.

Ranking

You can find your standings among all other players in the menu under *Statistics and Data* >> *Users* >> *User Ranking* on the TechForX webpage. Here you can also infer how active you are compared to other traders.

M.1.3 Classification of time horizons

Table A1 shows the descriptive statistics of the respective Delphi results. As described in Section 3.3, Delphi results were aggregated to four time horizons per thesis. For five theses, this resulted in N=5 observations per time horizon. Assuming one has no clue about in which time horizon a particular thesis is more likely to come true, one would presumably assign a probability of 25% for each time horizon. This would reflect complete uncertainty about the outcome.

Table A1: Descriptive statistics of the Delphi results

Time Horizon	Number of observations	Mean Delphi result	Standard deviation	Difficulty to predict
Up to 2010	5	14.4	6.1	Moderate
2011-2020	5	49.5	9.4	Easy
Later	5	17.9	12.2	Hard
Never	5	18.2	15.7	Hard

In our case, the mean Delphi results in Table A1 reveal large differences in how Delphi participants assessed the likelihood of occurrence of the theses for the different time horizons. For example, on average, the five theses were judged to come true 2011-2020 with a probability of 49.5% (standard deviation: 9.4). The absolute deviation from maximum uncertainty (i.e. 25%) was 24.5 percentage points – by far the largest for all time horizons. This indicates that it might have been relatively easy to anticipate this time horizon as more likely for the theses to come true. Thus, time horizon '2011-2020' was considered as an easy-to-assess environment. On the contrary, the time horizons 'Never' and 'Later' were classified as hard-to-assess environments. First, the mean Delphi results for these time horizons showed the lowest absolute deviation from maximum uncertainty with 6.8 and 7.1 percentage points, respectively. Second, their standard deviations from the mean (15.7 and 12.2) were relatively large. This indicates high variations in the Delphi results for these particular time horizons, which indicates uncertainty about the outcome. Finally, showing an absolute deviation from maximum uncertainty of still above 10 percentage points and the lowest standard deviation from the mean Delphi result, the time horizon 'Up to 2010' was classified as a moderate-to-assess environment..

M.2 Appendix to the Laboratory Experiment

M.2.1 Instructions (NGT)

ESTIMATE-TALK-ESTIMATE - ORAL CONSENT SCRIPT AND INSTRUCTIONS

You are invited to take part in a research study. The purpose of this study is to analyze how information is aggregated from groups.

During the process, **you are not allowed to use external sources of information**. But you are allowed to use pen and paper for your notes. Use the predefined sheets we distribute and label them with your LabID. We ask you to submit these sheets at the end of the study.

You will participate in a process called **estimate-talk-estimate**. Participation will be conducted in three steps:

1. You will work as an individual, providing estimates for 10 questions on a predefined questionnaire. You have **10 minutes** for this task.
2. You will participate in a group meeting.
    a. The goal of this meeting is to discuss the answers of the 10 questions with your group members.
    b. Your group has to fill out a questionnaire with your achieved group answers for the 10 questions.
    c. For every question, note the achieved group answer. You will need it for step 3.

    You have **20 minutes** for group discussion.

3. You will again work as an individual. As a part of the questionnaire in step 3, **you will provide your final individual estimates** for the 10 questions.

At the end of the process, the **final group answer** for each of the 10 questions will be calculated as the **median of your final individual estimates**.

Depending on your group performance, you can win money!

We calculate your group performance using the median of the error ratios for each question. The error ratio is defined by

- If group answer > correct answer ➔ error ratio = group answer / correct answer
- If group answer < correct answer ➔ error ratio = correct answer / group answer

Basically, **the closer your group answers are to the correct answers**, the lower the median error ratios, and **the better your group will perform!**

We calculate a ranking of the median error ratios of the 12 meeting groups in this experiment series. The best 6 groups will be remunerated: **$50 for 1st and 2nd; $25 for 3rd and 4th; $15 for 5th and 6th.** The money will be split equally among group members. In case groups achieve the same score, the money will be split among groups.

To be able to receive the additional pay-offs, make sure you **label all your materials with your LabID** so that we can contact you after we have determined the winners!

There should be no risk and discomfort in this task. You do not have to agree to be in this study. Participation is completely voluntary. If you have any questions about the study, please contact us via graefe@itas.fzk.de

If you have any questions about your rights as a research participant, or if you think you have not been treated fairly, you may call the University of Pennsylvania Institutional Review Board (IRB) at 215-898-2614.

M.2.2 Revealing prior individual estimates (NGT)

STEP 1: Individual task

LabID: _____

1. For each question,
   a. For each question, write down your individual estimates
   b. Check in the table how confident you are in your individual estimates from 1 to 7 (1: not confident at all; 7: extremely confident)

No	Question	Individual estimate	Confidence in accuracy of individual estimate						
			Not at all confident						Extremely confident
			1	2	3	4	5	6	7
1	In 1900, the percentage of the total US population that was aged 65 and over was 4.1 %. What was the percentage in 2000?								
2	In 2006, the US share of worldwide Internet users was 17.3%. What was China's share of worldwide Internet users in 2006?								
3	In 1970, 6.9% of the gross domestic product (GDP) was spent on health expenditures. What was the percentage spent in 2003?								
4	In 1980, 17.0% of the US population (25 years and over) had completed 4 years of college or more. What was the percentage in 2006?								
5	In 2000, 13.3% of the world's population lived in Africa. What was the percentage living in Australia (incl. Oceania) in 2000?								
6	In 1974, 4.0% of US children between 6 and 11 years old were overweight. What was the percentage in 2004?								
7	In 1970, the percentage of all US children (under 18 years) who lived with two parents was 85.2%. What was the percentage in 2005?								
8	In 1994, the percentage of all US households that had a personal computer was 24.1%. What was the percentage in 2002?								
9	In 1950, the US share of the world's motor vehicle production was 75.7%. What was the US share in 2006?								
10	In 1960, the percentage of the total US population that was aged under 5 years was 11.3%. What was the percentage in 2000?								

M.2.3 Revealing group estimates (NGT and FTF)

STEP 2 - GROUP

For each question,

2. Write down the achieved group answer
3. Check in the table how confident you as a group are in the achieved group answer from 1 to 7 (1: not confident at all; 7: extremely confident)

No	Question	Achieved group estimate	Confidence in accuracy of achieved group estimate						
			Not at all confident						Extremely confident
			1	2	3	4	5	6	7
1	In 1900, the percentage of the total US population that was aged 65 and over was 4.1 %. What was the percentage in 2000?								
2	In 2006, the US share of worldwide Internet users was 17.3%. What was China's share of worldwide Internet users in 2006?								
3	In 1970, 6.9% of the gross domestic product (GDP) was spent on health expenditures. What was the percentage spent in 2003?								
4	In 1980, 17.0% of the US population (25 years and over) had completed 4 years of college or more. What was the percentage in 2006?								
5	In 2000, 13.3% of the world's population lived in Africa. What was the percentage living in Australia (incl. Oceania) in 2000?								
6	In 1974, 4.0% of US children between 6 and 11 years old were overweight. What was the percentage in 2004?								
7	In 1970, the percentage of all US children (under 18 years) who lived with two parents was 85.2%. What was the percentage in 2005?								
8	In 1994, the percentage of all US households that had a personal computer was 24.1%. What was the percentage in 2002?								
9	In 1950, the US share of the world's motor vehicle production was 75.7%. What was the US share in 2006?								
10	In 1960, the percentage of the total US population that was aged under 5 years was 11.3%. What was the percentage in 2000?								

M.2.4 Revealing posterior estimates and ex post evaluation (NGT)

STEP 3: Individual task

I. For each question,

1. Again, write down the achieved group answers from your group discussion. Just copy them from your individual questionnaire before.
2. Check in the table how confident you **as an individual** are in the achieved group answers from the previous discussion (1: not confident at all; 7: extremely confident)

No	Question	Achieved group estimate	Confidence in accuracy of achieved group estimate						
			Not at all confident						Extremely confident
			1	2	3	4	5	6	7
1	In 1900, the percentage of the total US population that was aged 65 and over was 4.1 %. What was the percentage in 2000?								
2	In 2006, the US share of worldwide Internet users was 17.3%. What was China's share of worldwide Internet users in 2006?								
3	In 1970, 6.9% of the gross domestic product (GDP) was spent on health expenditures. What was the percentage spent in 2003?								
4	In 1980, 17.0% of the US population (25 years and over) had completed 4 years of college or more. What was the percentage in 2006?								
5	In 2000, 13.3% of the world's population lived in Africa. What was the percentage living in Australia (incl. Oceania) in 2000?								
6	In 1974, 4.0% of US children between 6 and 11 years old were overweight. What was the percentage in 2004?								
7	In 1970, the percentage of all US children (under 18 years) who lived with two parents was 85.2%. What was the percentage in 2005?								
8	In 1994, the percentage of all US households that had a personal computer was 24.1%. What was the percentage in 2002?								
9	In 1950, the US share of the world's motor vehicle production was 75.7%. What was the US share in 2006?								
10	In 1960, the percentage of the total US population that was aged under 5 years was 11.3%. What was the percentage in 2000?								

II. Now, for each question,

1. Again, rate how you would now, after the meeting, individually estimate each question.
2. Check in the table how confident you are in your individual ex post estimate from 1 to 7 (1: not confident at all; 7: extremely confident)

The final group answers will be calculated as the average

of the final individual estimates of all group members

Remember, based on the accuracy of the final group answers, you can win money.

[Your confidence scores do not influence the pay-off]

No	Question	Individual ex post estimate	Confidence in individual ex post estimate						
			Not at all confident						Extremely confident
			1	2	3	4	5	6	7
1	In 1900, the percentage of the total US population that was aged 65 and over was 4.1 %. What was the percentage in 2000?								
2	In 2006, the US share of worldwide Internet users was 17.3%. What was China's share of worldwide Internet users in 2006?								
3	In 1970, 6.9% of the gross domestic product (GDP) was spent on health expenditures. What was the percentage spent in 2003?								
4	In 1980, 17.0% of the US population (25 years and over) had completed 4 years of college or more. What was the percentage in 2006?								
5	In 2000, 13.3% of the world's population lived in Africa. What was the percentage living in Australia (incl. Oceania) in 2000?								
6	In 1974, 4.0% of US children between 6 and 11 years old were overweight. What was the percentage in 2004?								
7	In 1970, the percentage of all US children (under 18 years) who lived with two parents was 85.2%. What was the percentage in 2005?								
8	In 1994, the percentage of all US households that had a personal computer was 24.1%. What was the percentage in 2002?								
9	In 1950, the US share of the world's motor vehicle production was 75.7%. What was the US share in 2006?								
10	In 1960, the percentage of the total US population that was aged under 5 years was 11.3%. What was the percentage in 2000?								

III. Rate the following two questions, again from 1 (very low) to 7 (very high)

Task	Very low 1	2	3	4	5	6	Very high 7
Rate your group discussion overall on							
- Cooperation							
- Overall accuracy							
- Disagreement about answers							
- Confidence in group answers							
- Conflict							
Rate yourself overall							
- on your performance in the meeting							
- on your overall accuracy of your individual estimates							
How accurate do you think will the final group results of the estimate-talk-estimate process (calculated as the average of the final individual estimates within your group) be?							
To what extent did you feel free to participate and contribute your opinion in the estimate-talk-estimate process?							
To what extent did you feel your time was well spent in participating in the estimate-talk-estimate process?							
How satisfied are you with the estimate-talk-estimate process overall?							
How difficult was it to participate in the estimate-talk-estimate process?							
To what extent do you feel estimate-talk-estimate processes are an effective way to deal with the problem?							

Any further comments on the study?

_____

_____

_____

THANKS FOR YOUR PARTICIPATION!

Technical Appendix

T.1 Calculation of Error Measure

This Appendix explains the calculation of the absolute percentage error, which is the error measure used for the analyses in this work. For an empirical comparison of different error measures and recommendations for when to use certain measures see Armstrong and Collopy (1992).

The following notation is used for the definitions of error measures that follow:

- m the forecasting method,
- h the horizon being forecasted
- s the series being forecasted
- $F_{m,h,s}$ the forecast from method m for horizon h of series s
- $A_{h,s}$ the actual value at horizon h of series s
- S the number of series being summarized

The absolute percentage error (APE) for a particular forecasting method for a given horizon of a particular series is defined as

$$APE_{m,h,s} = \left| \frac{F_{m,h,s} - A_{h,s}}{A_{h,s}} \right|$$

The mean absolute percentage error (MAPE) is defined as

$$MAPE_{m,h} = \frac{\sum_{s-1}^{S} APE_{m,h,s}}{S} * 100$$

The median absolute percentage error (MdAPE) is defined as $MdAPE_{m,h}$= Observation $\frac{S+1}{2}$ if S is odd, or the mean of observations $\frac{S}{2}$ and $\frac{S}{2}+1$ if S is even, where the observations are ranked-ordered by $APE_{m,h,s}$.

T.2 Remarks to Statistical Analyses

The data presented in the Chapters 4 to 8 was analyzed using tests of statistical significance. All statistics were calculated using SPSS 15.0.

Following the recommendations of Armstrong (2007), it is suggested that readers focus on effect sizes in the data. In being reported in footnotes, readers can skip tests of statistical significance if they like. Significance levels are indicated in tables using asterisks. The following coding was used to indicate levels of significance:

* ➔ p-value ≤ 0.05

** ➔ p-value ≤ 0.01

For full disclosure, all data that has been analyzed in this study can be found online. The links to the data are provided in the respective chapters or can be found in the supporting online material in the Appendix. For the calculation of the error measures used in this work, see the Technical Appendix T.1.

Supporting Online Material

For the purpose of full disclosure, all data that has been analyzed in this work is available online. This Appendix provides the links to the respective data files for the analyses in the various chapters.

Chapter 4: Validity of Prediction Markets for Long-term Forecasting

http://spreadsheets.google.com/pub?key=pr1ZdfEZ874kIU-F05RnsIg

Chapter 5: The Value of Experts in Prediction Markets

http://spreadsheets.google.com/pub?key=pr1ZdfEZ874kJrmlGDHLvJA

Chapter 6: Relative Accuracy of Prediction Markets on a Quantitative Judgment Task

http://spreadsheets.google.com/pub?key=pr1ZdfEZ874nHtHGW_CT7dA

Chapter 7: Perceptions of Prediction Markets

http://spreadsheets.google.com/pub?key=pr1ZdfEZ874nHtHGW_CT7dA

Chapter 8: Advice-taking from Prediction Markets, Meetings, and the Delphi Method

http://andreas-graefe.org/data/advicetaking.sav

Furthermore, the introductory audio tutorials for Delphi and prediction market groups in the laboratory experiment, described in Section 3.4, can be found online:

- **Delphi tutorial**: http://andreas-graefe.org/data/Delphi_Tutorial.ppsx
- **Prediction market tutorial**: http://andreas-graefe.org/data/pm_tutorial.pps

Bibliography

ARKES, H. R. & HARKNESS, A. R. (1980). The Effect of Making a Diagnosis on the Subsequent Recognition of Symptoms, *Journal of Experimental Psychology: Human Learning and Memory*, 6, 568-575.

ARMSTRONG, J. S. (1980). The Seer-Sucker Theory: The Value of Experts in Forecasting, *Technology Review*, 83, 16-24.

ARMSTRONG, J. S. (2001). Combining Forecasts. In: J. S. Armstrong (Eds.), *Principles of Forecasting. A Handbook for Researchers and Practitioners.* Norwell; Kluwer Academic Publishers, pp. 417-439.

ARMSTRONG, J. S. (2006). How to Make Better Forecasts and Decisions: Avoid Face-to-Face Meetings, *Foresight*, 5, 3-8.

ARMSTRONG, J. S. (2007). Significance Tests Harm Progress in Forecasting, *International Journal of Forecasting*, 23, 321-327.

ARMSTRONG, J. S. & COLLOPY, F. (1992). Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons, *International Journal of Forecasting*, 8, 69-80.

BERG, J., NELSON, F. & RIETZ, T. A. (2008a). Prediction Market Accuracy in the Long Run, *International Journal of Forecasting*, 24, 285-300.

BERG, J., NELSON, F. D., NEUMANN, G. R. & RIETZ, T. (2008b). *Was There Any Surprise About Obama's Election?*, Available at http://www.forecastingprinciples.com/PM/images/articles/berg_obama_surprise.pdf.

BEST, R. J. (1974). An Experiment in Delphi Estimation in Marketing Decision Making, *Journal of Marketing Research*, 11, 448-452.

BOJE, D. M. & MURNIGHAN, J. K. (1982). Group Confidence Pressures in Iterative Decisions, *Management Science*, 28, 1187-1196.

BONACCIO, S. & DALAL, R. S. (2006). Advice Taking and Decision-Making: An Integrative Literature Review, and Implications for the Organizational Sciences, *Organizational Behavior and Human Decision Processes*, 101, 127-151.

BONNER, S. E., HASTIE, R., SPRINKLE, G. B. & YOUNG, S. M. (2000). A Review of the Effects of Financial Incentives on Performance in Laboratory Tasks: Implications for Management Accounting, *Journal of Management Accounting Research*, 12, 19-64.

BROCKHOFF, K. (1975). The Performance of Forecasting Groups in Computer Dialogue and Face-to-Face Discussion. In: H. A. Linstone & Turoff, M. (Eds.), *The Delphi Method: Techniques and Applications.* London; Addison-Wesley, pp. 291-322.

CABALLE, J. & SAKOVICS, J. (2003). Speculating against an Overconfident Market, *Journal of Financial Markets*, 6, 199-225.

CAIN, M. & DRAKOS, N. (2008). *Prediction Markets*. In: Hype Cycle for Social Software, Gartner Research, Available at http://blog.kwiqq.com/wp-content/uploads/2009/01/hype_cycle_for_social_softwa_158239.pdf.

CAMERER, C. (1998). Can Asset Markets Be Manipulated? A Field Experiment with Racetrack Betting, *Journal of Political Economy*, 106, 457-482.

CHEN, K.-Y., FINE, L. R. & HUBERMAN, B. A. (2004). Eliminating Public Knowledge Biases in Information-Aggregation Mechanisms, *Management Science*, 50, 983-994.

CHEN, K.-Y. & PLOTT, C. R. (2002). *Information Aggregation Mechanisms: Concept, Design and Implementation for a Sales Forecasting Problem*. In: Social Science Working Paper No.1131, California Institute of Technology, Pasadena. Available at http://www.hss.caltech.edu/SSPapers/wp1131.pdf.

CHERRY, S. (2007). Bet on It! Can a Stock Market of Ideas Help Companies Predict the Future?, *IEEE Spectrum*, 44, 48-53.

CHRISTIANSEN, J. D. (2007). Prediction Markets: Practical Experiments in Small Markets and Behaviours Observed, *Journal of Prediction Markets*, 1, 17-41.

COWGILL, B., WOLFERS, J. & ZITZEWITZ, E. (2008). *Using Prediction Markets to Track Information Flows: Evidence from Google*, Working Paper. Available at http://bocowgill.com/GooglePredictionMarketPaper.pdf.

CUHLS, K. (2003). From Forecasting to Foresight Processes - New Participative Foresight Activities in Germany, *Journal of Forecasting*, 22, 93-111.

DAHAN, E., LO, A. W., POGGIO, T., CHAN, N. & KIM, A. (2007). *Securities Trading of Concepts (Stoc)*, Working Paper. Available at http://www.anderson.ucla.edu/faculty/ely.dahan/content/chan_dahan_lopoggio.pdf.

DALKEY, N. & HELMER, O. (1963). An Experimental Application of the Delphi Method to the Use of Experts, *Management Science*, 9, 458-467.

DIETZ, T. (1987). Methods for Analyzing Data from Delphi Panels: Some Evidence from a Forecasting Study, *Technological Forecasting and Social Change*, 31, 79-85.

EINHORN, H. J. & HOGARTH, R. M. (1978). Confidence in Judgment: Persistence of the Illusion of Validity, *Psychological Review*, 85, 395-416.

ERIKSON, R. S. & WLEZIEN, C. (2008). Are Political Markets Really Superior to Polls as Election Predictors? Forthcoming in *Public Opinion Quarterly*. Available at http://poq.oxfordjournals.org/cgi/content/summary/nfn010v1.

FENN, J. & LINDEN, A. (2005). *Gartner's Hype Cycle Special Report for 2005*, Gartner Research, Available at http://www.gartner.com/resources/130100/130115/gartners_hype_c.pdf.

FISCHER, G. W. (1981). When Oracles Fail--a Comparison of Four Procedures for Aggregating Subjective Probability Forecasts, *Organizational Behavior and Human Performance*, 28, 96-110.

FORSYTHE, R., NELSON, F., NEUMANN, G. R. & WRIGHT, J. (1992). Anatomy of an Experimental Political Stock Market, *American Economic Review*, 82, 1142-1142.

FORSYTHE, R., RIETZ, T. A. & ROSS, T. W. (1999). Wishes, Expectations and Actions: A Survey on Price Formation in Election Stock Markets *Journal of Economic Behavior & Organization*, 39, 83-110.

FRIEDEWALD, M., VON OERTZEN, J. & CUHLS, K. (2007). *European Perspectives on the Information Society (Epis). Delphi Report*. In: ETEPS - European Techno-Economic Policy Support Network, Fraunhofer ISI, Karlsruhe, Germany. Available at http://epis.jrc.es/documents/Deliverables/EPIS%202-3-1%20Delphi%20Report.pdf.

GLASER, M. & WEBER, M. (2007). Overconfidence and Trading Volume, *Geneva Risk and Insurance Review*, 32, 1-36.

GLASS, G. V., PECKHAM, P. D. & SANDERS, J. R. (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance, *Review of Educational Research*, 42, 237-288.

GORDON, T. & PEASE, A. (2006). Rt Delphi: An Efficient, "Round-Less" Almost Real Time Delphi Method, *Technological Forecasting and Social Change*, 73, 321-333.

GRAEFE, A. (2006). Information Markets as a Participatory Method for Technology Assessment. Presentation at the *NTA2 - Second Conference of the 'TA Network': Technology Assessment in World Society*, Berlin, Germany, November 22-24.

GRAEFE, A. (2007a). Aspects of Information Aggregation - a Comparison of the Delphi Method and Prediction Markets. Presentation at the *27th International Symposium on Forecasting*, New York City, USA, June 24-27.

GRAEFE, A. (2007b). Folgenabschätzung Durch Prognosemärkte, *Technikfolgenabschätzung - Theorie und Praxis*, 16, 66-73.

GRAEFE, A. (2007c). Forecasting Mit Prognosemärkten. In: A. Bora et al. (Eds.), *Technology Assessment in Der Weltgesellschaft.* Berlin; Edition Sigma, pp. 439-444.

GRAEFE, A. (2007d). Prediction Markets for Participatory Decision-Making. Presentation at the *Third International Conference on Technology, Knowledge, and Society*, Cambridge, UK, January 9-12.

GRAEFE, A. (2007e). Using Markets to Forecast the Future. Presentation at the *Third International Conference on Organisational Foresight*, Glasgow, UK, August 16-18.

GRAEFE, A. (2008a). Can You Beat the Market? Accuracy of Individual and Group Post-Prediction Market Judgments. Presentation at the *Third Workshop on Prediction Markets*, Chicago, USA, July 9.

GRAEFE, A. (2008b). Group Decision Making - Meetings, Nominal Groups, Delphi and Prediction Markets Compared. Presentation at the *28th International Symposium on Forecasting*, Nice, France, June 22-25.

GRAEFE, A. (2008c). Group Decision Making - Meetings, Nominal Groups, Delphi, and Prediction Markets Compared. Presentation at the *Third Workshop on Prediction Markets*, Chicago, USA, July 9.

GRAEFE, A. (2008d). Harnessing Collective Knowledge - Use Markets to Improve Your Forecasts. Presentation at the *Forecasting Summit*, Boston, USA, September 17.

GRAEFE, A. (2008e). Prediction Markets – Defining Events and Motivating Participation, *Foresight - The International Journal of Applied Forecasting*, 2008 (Spring), 30-32.

GRAEFE, A. (2008f). Accuracy of Individual and Group Post-Prediction Market Judgments. Presentation at the INFORMS Annual Meeting, Washington D.C., October 15.

GRAEFE, A. & ARMSTRONG, J. S. (2008a). *Advice-Taking from Prediction Markets, Meetings, and the Delphi Method*, Working Paper. Available at http://andreas-graefe.org/images/articles/graefe_armstrong_advice_taking.pdf.

GRAEFE, A. & ARMSTRONG, J. S. (2008b). *Comparing Face-to-Face Meetings, Nominal Groups, Delphi and Prediction Markets on an Estimation Task*, Working paper. Available at http://www.forecastingprinciples.com/PM/images/articles/graefe_armstrong_GDM_methods.pdf.

GRAEFE, A., ARMSTRONG, J. S., CUZÁN, A. G. & JONES JR, R. J. (2009a). Combined Forecasts of the 2008 Election: The Pollyvote, *Foresight - The International Journal of Applied Forecasting*, 2009 (Spring), 41-42.

GRAEFE, A., ARMSTRONG, J. S. & GREEN, K. C. (2009b). *Using Prediction Markets to Solve Complex Problems - an Application to the 'Climate Bet'*, Wharton School, University of Pennsylvania, Working Paper. Available at http://andreas-graefe.org/images/articles/CBPM.pdf.

GRAEFE, A., LUCKNER, S. & WEINHARDT, C. (2009c). Prediction Markets - a Toolkit for Foresight. Forthcoming in *Futures*. Available at http://andreas-graefe.org/images/articles/PM_Foresight.pdf.

GRAEFE, A. & ORWAT, C. (2007). Prediction Markets as a Mechanism for Public Engagement? A First Classification and Open Questions, *International Journal of Technology, Knowledge and Society*, 3, 137-142.

GRAEFE, A. & WEINHARDT, C. (2008). Long-Term Forecasting with Prediction Markets - a Field Experiment on Applicability and Expert Confidence, *Journal of Prediction Markets*, 2, 71-92.

GREEN, K. C., ARMSTRONG, J. S. & GRAEFE, A. (2007). Methods to Elicit Forecasts from Groups: Delphi and Prediction Markets Compared, *Foresight - The International Journal of Applied Forecasting*, 2007 (Fall), 17-20.

GUSTAFSON, D. H., SHUKLA, R. K., DELBECQ, A. & WALSTER, G. W. (1973). A Comparative Study of Differences in Subjective Likelihood Estimates Made by Individuals, Interacting Groups, Delphi Groups, and Nominal Groups, *Organizational Behavior and Human Performance*, 9, 280–291.

HANSEN, J., SCHMIDT, C. & STROBEL, M. (2004). Manipulation in Political Stock Markets - Preconditions and Evidence, *Applied Economics Letters*, 11, 459-463.

HANSON, R. (2003). Combinatorial Information Market Design, *Information Systems Frontiers*, 5, 107-119.

HANSON, R. (2006). Designing Real Terrorism Futures, *Public Choice*, 128, 1-18.

HANSON, R. (2007). The Policy Analysis Market - a Thwarted Experiment in the Use of Prediction Markets for Public Policy, *Innovations: Technology, Governance, Globalization (MIT Press)*, 2, 73-88.

HANSON, R., OPREA, R. & PORTER, D. (2006). Information Aggregation and Manipulation in an Experimental Market, *Journal of Economic Behavior & Organization*, 60, 449-459.

HARVEY, N. & FISCHER, I. (1997). Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility, *Organizational Behavior and Human Decision Processes*, 70, 117-133.

HAYEK, F. A. (1945). The Use of Knowledge in Society, *American Economic Review*, 35, 519-530.

HENRY, R. A. (1993). Group Judgment Accuracy: Reliability and Validity of Postdiscussion Confidence Judgments, *Organizational Behavior and Human Decision Processes*, 56, 11-27.

HENRY, R. A. (1995). Using Relative Confidence Judgments to Evaluate Group Effectiveness, *Basic & Applied Social Psychology*, 16, 333-350.

HENRY, R. A. & SNIEZEK, J. A. (1993). Situational Factors Affecting Judgments of Future Performance, *Organizational Behavior and Human Decision Processes*, 54, 104-132.

JANIS, I. (1972). *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*, Boston: Houghton Mifflin.

JANSSEN, S. (2007). *The World Almanac and Book of Facts 2008*, New York City: World Almanac Education Group Inc.

JONES, R. J. (2008). The State of Presidential Election Forecasting: The 2004 Experience, *International Journal of Forecasting*, 24, 310-321.

KELLEY, H. H. & THIBAUT, J. W. (1954). Experimental Studies of Group Problem Solving and Process. In: G. Lindzey (Eds.), *Handbook of Social Psychology: Special Fields and Applications.* Reading, MA; Addison-Wesley, pp. 735-785.

KING, R. (2006). Workers, Place Your Bets, *BusinessWeek*, August 3, Available at http://www.businessweek.com/technology/content/aug2006/tc20060803_012437.htm.

KOEHLER, D. J., BRENNER, L. & GRIFFIN, D. (2002). The Calibration of Expert Judgment: Heuristics and Biases Beyond the Laboratory. In: T. Gilovich et al. (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge University Press, pp. 489–509.

LARRECHE, J. C. & MOINPOUR, R. (1983). Managerial Judgment in Marketing: The Concept of Expertise, *Journal of Marketing Research*, 20, 110-121.

LARRICK, R. P. & SOLL, J. B. (2006). Intuitions About Combining Opinions: Misappreciation of the Averaging Principle, *Management Science*, 52, 111-127.

LAUGHLIN, P. R. (1999). Collective Induction: Twelve Postulates, *Organizational Behavior and Human Decision Processes*, 80, 50-69.

LUCKNER, S. (2008). *Predictive Power of Markets - Prediction Accuracy, Incentive Schemes, and Traders' Biases*, Doctoral dissertation, School of Economics and Business Engineering, Universität Karlsruhe (TH).

LUCKNER, S., SCHRÖDER, J. & SLAMKA, C. (2008). On the Forecast Accuracy of Sports Prediction Markets In: H. Gimpel et al. (Eds.), *Negotiation, Auctions & Market Engineering, Lecture Notes in Business Information Processing (Lnbip).* Springer-Verlag, pp. 227-234.

LUCKNER, S. & WEINHARDT, C. (2007). How to Pay Traders in Information Markets? Results from a Field Experiment, *Journal of Prediction Markets*, 1, 147-156.

MAIER, N. R. F. & HOFFMAN, L. R. (1960). Quality of First and Second Solutions in Group Problem Solving, *Journal of Applied Psychology*, 44, 310-323.

MANYIKA, J., ROBERTS, R. P. & SPRAGUE, K. L. (2007). *Eight Business Technology Trends to Watch*, McKinsey Quarterly. Available at http://www.mckinseyquarterly.com/article_page.aspx?ar=2080&l2=13&l3=11.

MATEOS-GARCIA, J., GEUNA, A. & STEINMUELLER, E. W. (2007). *Discussion Paper on the State of the Art of the European Creative Content Industry and Market & National/Industrial Initiatives*. In: European Perspectives on Information Society, Institute for Prospective Technological Studies, Available at http://epis.jrc.ec.europa.eu/documents/Deliverables/DP1%20Final%20version%201.pdf.

MCGRATH, J. E. (1984). *Groups: Interaction and Performance*, Prentice-Hall, Englewood Cliffs, NJ.

MURPHY, M. K., BLACK, N. A., LAMPING, D. L., MCKEE, C. M., SANDERSON, C. F. B., ASKHAM, J. & MARTEAU, T. (1998). Consensus Development Methods, and Their Use in Clinical Guideline Development, *Health Technology Assessment*, 2, 1-88.

ODEAN, T. (1998). Volume, Volatility, Price, and Profit When All Traders Are above Average, *The Journal of Finance*, 53, 1887-1934.

ORTNER, G. (1998). *Forecasting Markets - an Industrial Application: Part I*, TU Vienna, Working Paper. Available at http://www.imw.tuwien.ac.at/apsm/fmaia2.pdf.

PENNOCK, D. M. (2004). *A Dynamic Pari-Mutuel Market for Hedging, Wagering, and Information Aggregation*, ACM Conference on Electronic Commerce, New York, NY, USA, May 17-20.

PENNOCK, D. M., GILES, C. L. & NIELSEN, F. A. (2001a). The Real Power of Artificial Markets, *Science*, 291, 987-988.

PENNOCK, D. M., LAWRENCE, S., GILES, C. L. & NIELSEN, F. A. (2001b). *The Power of Play: Efficiency and Forecast Accuracy of Web Market Games*. In: Technical Report 2000-168, NEC Research Institute, Available at http://artificialmarkets.com/am/pennock-neci-tr-2000-168.pdf.

REMUS, W., O'CONNOR, M. & GRIGGS, K. (1998). The Impact of Incentives on the Accuracy of Subjects in Judgmental Forecasting Experiments, *International Journal of Forecasting*, 14, 515-522.

RHODE, P. W. & STRUMPF, K. S. (2004). Historical Presidential Betting Markets, *Journal of Economic Perspectives*, 18, 127-141.

RIDER, P. R. (1929). On the Distribution of the Ratio of Mean to Standard Deviation in Small Samples from Non-Normal Universes, *Biometrika*, 21, 124-143.

RIETZ, T. A. (2005). *Behavioral Mis-Pricing and Arbitrage in Experimental Asset Markets*, University of Iowa, Working Paper. Available at http://www.biz.uiowa.edu/faculty/trietz/papers/market.pdf.

ROONEY, B. (2008). Oddsmakers See Bailout Ok Soon. *CNNMoney.com*, September 24, Available at http://money.cnn.com/2008/09/24/news/economy /intrade_bailout/?postversion=2008092415

ROSENBLOOM, E. S. & NOTZ, W. (2006). Statistical Tests of Real-Money Versus Play-Money Prediction Markets, *Electronic Markets*, 16, 63-69.

ROWE, G. & WRIGHT, G. (1996). The Impact of Task Characteristics on the Performance of Structured Group Forecasting Techniques, *International Journal of Forecasting*, 12, 73-89.

ROWE, G. & WRIGHT, G. (1999). The Delphi Technique as a Forecasting Tool: Issues and Analysis, *International Journal of Forecasting*, 15, 353-375.

ROWE, G. & WRIGHT, G. (2001). Expert Opinions in Forecasting: The Role of the Delphi Technique. In: J. S. Armstrong (Eds.), *Principles of Forecasting - a Handbook for Researchers and Practitioners.* Boston, MA; Kluwer Academic Publishers, pp. 125-144.

SERVAN-SCHREIBER, E., WOLFERS, J., PENNOCK, D. M. & GALEBACH, B. (2004). Prediction Markets: Does Money Matter?, *Electronic Markets*, 14, 243 - 251.

SNIEZEK, J. A. & HENRY, R. A. (1990). Revision, Weighting, and Commitment in Consensus Group Judgment, *Organizational Behavior and Human Decision Processes*, 45, 66-84.

SOUKHOROUKOVA, A. (2007). *Produktinnovation Mit Informationsmärkten*, Doctoral dissertation, Universität Passau.

SPANN, M., ERNST, H., SKIERA, B. & SOLL, J. H. (2007). Identification of Lead Users for Consumer Products Via Virtual Stock Markets, Forthcoming in *Journal of Product Innovation Management.*

SPANN, M. & SKIERA, B. (2003). Internet-Based Virtual Stock Markets for Business Forecasting, *Management Science*, 49, 1310-1326.

SPANN, M. & SKIERA, B. (2009). Sports Forecasting: A Comparison of the Forecast Accuracy of Prediction Markets, Betting Odds and Tipsters, *Journal of Forecasting*, 28, 55-72.

STIX, G. (2008). When Markets Beat the Polls, *Scientific American Magazine*, 298, 38-45.

SUROWIECKI, J. (2004). *The Wisdom of Crowds. Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*, Little, Brown Book Group.

TETLOCK, P. C. (2008). *Liquidity and Prediction Market Efficiency*, SSRN working paper. Available at http://ssrn.com/paper=929916.

TETLOCK, P. E. (2006). *Expert Political Judgment: How Good Is It? How Can We Know?*, Princeton University Press.

THALER, R. H. & ZIEMBA, W. T. (1988). Anomalies: Parimutuel Betting Markets: Racetracks and Lotteries, *The Journal of Economic Perspectives*, 2, 161-174.

TVERSKY, A. & KAHNEMAN, D. (1974). Judgment under Uncertainty: Heuristics and Biases, *Science*, 185, 1124-1131.

TZIRALIS, G. & TATSIOPOULOS, I. (2007). Prediction Markets; an Extended Literature Review, *Journal of Prediction Markets*, 1, 75-91.

VAN DE VEN, A. H. & DELBECQ, A. L. (1971). Nominal Versus Interacting Group Processes for Committee Decision Making Effectiveness, *Academy of Management Journal*, 14, 203-212.

VAN DE VEN, A. H. & DELBECQ, A. L. (1974). The Effectiveness of Nominal, Delphi, and Interacting Group Decision Making Processes, *Academy of Management Journal*, 17, 605-621.

WEINHARDT, C., HOLTMANN, C. & NEUMANN, D. (2003). Market-Engineering, *Wirtschaftsinformatik*, 45, 635-640.

WOLFERS, J. & ZITZEWITZ, E. (2004). Prediction Markets, *Journal of Economic Perspectives*, 18, 107-126.

WOUDENBERG, F. (1991). An Evaluation of Delphi, *Technological Forecasting and Social Change*, 40, 131-150.

YANIV, I. (2004a). The Benefit of Additional Opinions, *Current Directions in Psychological Science*, 13, 75-78.

YANIV, I. (2004b). Receiving Other People's Advice: Influence and Benefit, *Organizational Behavior and Human Decision Processes*, 93, 1-13.

YANIV, I. & KLEINBERGER, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation, *Organizational Behavior and Human Decision Processes*, 83, 260-281.

Author Index

W

Y

Z