Engineering Delphi-Markets for crowd-based Prediction
The FAZ.NET-Orakel and other cases

Zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der Fakultät für
Wirtschaftswissenschaften
am Karlsruher Institut für Technologie (KIT)

genehmigte
DISSERTATION

von
Simon Andreas Kloker (M.Sc.)

Tag der mündlichen Prüfung: 31. Oktober 2018
Referent: Prof. Dr. Christof Weinhardt
Korreferent: Prof. Dr. Rudi Studer

Karlsruhe, 2018

KARLSRUHE INSTITUTE OF TECHNOLOGY

Abstract

Department of Economics and Management
Institute of Information Systems and Marketing

Dr. rer. pol.

Engineering Delphi-Markets for crowd-based Prediction: The FAZ.NET-Orakel and other cases

by Simon Andreas KLOKER

Reliable forecasting is a key success factor of most organizations and companies. Where historical data is not available, the forecasts address questions in the far future, information is dispersed regarding location and form, or conflicting goals or values have to be considered, judgmental forecasting methods based on experts or the crowd are typically applied. However, several judgmental forecasting methods exist and each method has some individual weaknesses. Delphi-Markets are an integrated approach of prediction markets and Real-Time Delphi studies. Depending on their implementation, they allow to combine several properties of both approaches in order to overcome individual weaknesses. Three different ways to integrate the method are presented and discussed in this work. In order to better understand challenges and potentials of Delphi-Markets, the FAZ.NET-Orakel was instantiated and made publicly available for evaluation and improvement of an exemplary Delphi-Market under real-world conditions. In this context, four proposed improvements for the integrated approach were evaluated in four research projects. These projects correspond to the four sources of forecasting error according to the Judgmental Forecasting Improvement Model, introduced and derived in this dissertation as well. On the one hand, these improvements deal with common problems of prediction markets: Cognitive errors, such as partition dependence, and motivational errors, such as manipulation and fraud. On the other hand, these include common problems of Real-Time Delphi studies: The selection of experts for Delphi studies and retention during the surveys. As contributions to the overall IS research derived from the examinations of the Delphi-Markets and this dissertation, design principles for two extensions (social Real-Time Delphi and a crowd-based approach for manipulation and fraud detection) are formulated, implemented, tested, and suggested for application. Further, the role of complexity and expertise in the occurrence of the partition dependence bias is examined and a selection approach for experts for Delphi studies based on trading data is suggested and evaluated.

KARLSRUHE INSTITUTE OF TECHNOLOGY

Zusammenfassung

Fakultät für Wirtschaftswissenschaften
Institut für Informationswirschat und Marketing

Dr. rer. pol.

Engineering Delphi-Markets for crowd-based Prediction: The FAZ.NET-Orakel and other cases

von Simon Andreas KLOKER

Zuverlässige Prognosen sind ein wichtiger Erfolgsfaktor für die meisten Organisationen und Unternehmen. Wenn historische Daten nicht verfügbar sind, wenn sich die Prognosefragen auf die ferne Zukunft beziehen, wenn Informationen bezüglich ihrem Ort und ihrer Form verteilt sind und-oder wenn widersprüchliche Ziele oder Werte berücksichtigt werden müssen, wird dabei in der Regel auf Experten- oder "Crowd"-basierte Prognoseverfahren zurückgegriffen. Es gibt jedoch mehrere solcher Prognosemethoden die auf menschlicher Beurteilung basieren und jede dieser Methoden bringt individuelle Schwächen und Stärken mit sich. Delphi-Märkte sind ein integrierter Ansatz aus Prognosemärkten und Echtzeit-Delphi-Studien. Abhängig von ihrer Umsetzung erlauben sie es, mehrere Eigenschaften beider Ansätze zu kombinieren, um individuelle Schwächen zu überwinden. In dieser Arbeit werden drei verschiedene Wege zur Integration der Methoden vorgestellt und diskutiert. Um die Herausforderungen und Potenziale solcher Delphi-Märkte besser zu verstehen, zu verbessern und zu evaluieren, wurde im Rahmen dieser Arbeit das FAZ.NET-Orakel als beispielhafter Delphi-Markt instanziiert und öffentlich zugänglich gemacht. In diesem Zusammenhang wurden vier Verbesserungsvorschläge für den integrierten Ansatz in vier Forschungsprojekten evaluiert. Diese Projekte entsprechen den vier Arten für Prognosefehler nach dem "Judgmental Forecasting Improvement Model", welches auch in dieser Dissertation vorgestellt und abgeleitet wurde. Zum einen betreffen diese Verbesserungen häufig auftretende Probleme der Prognosemärkte: Fehler aufgrund von kognitiven Verzerrungen (z.B. "partition dependence bias") oder Fehler aufgrund von problematischer Motivation der Teilnehmer (z.B. Manipulation und Betrug). Zum anderen betreffen diese Verbesserungen häufige Probleme von Echtzeit-Delphi-Studien: Die Auswahl der Experten für Delphi-Studien und die Abwanderung der Teilnehmer während der Befragungen. Aus dieser Dissertation und der Untersuchung der Delphi-Märkte ergeben sich einige Beiträge zur Forschung im Bereich der Prognoseverfahren die auf menschlicher Beurteilung basieren, sowie anderen Fragestellungen aus dem Bereich der IS-Forschung. Dies sind unter anderem Design Prinzipien für zwei Erweiterungen der Delphi-Märkte (soziale Echtzeit-Delphi-Studien und ein Crowd-basierter Ansatz zur Manipulation- und Betrugserkennung), welche formuliert, implementiert, getestet und zur Anwendung vorgeschlagen werden. Weiterhin werden die Rolle von Komplexität und Expertise beim Auftreten des "partition dependence bias" untersucht und ein Ansatz zur Auswahl von Experten für Delphi-Studien auf Basis von Handelsdaten vorgeschlagen und bewertet.

Acknowledgements

First of all I would like to thank Prof. Dr. Christof Weinhardt for the opportunity to pursue my dissertation under his supervision. I would also like to thank my co-advisor Prof. Dr. Rudi Studer and the other members of the PhD committee, Prof. Dr. Hansjörg Fromm and Prof. Dr. Orestis Terzidis, for their time in reviewing the work at hand, for their helpful comments and the thorough and insightful discussions before, during and after the defense.

I have spent the last three years in an amazing team, although unfortunately its members changed much too quickly. For many of the friendships made here, I am deeply grateful. You pushed or carried me during that time, all in due course. The selection and naming of people would probably not do justice to the team at all. However, some colleagues have shared my doubts, thoughts and successes in a special way. Many thanks to Tim, Michael, Dominik, David, Christian, Benedikt, Greta, Florian and Tobi. For what? You probably know - thank you!

However, the energy that was burned during the course of writing this dissertation was replenished by much more streams. I would like to thank my wife Janny - always there to distract my mind from work towards more important things; always there when I need exchange or just proximity; always having the kind of "more" for me when I least expect it. I love you. Thanks apply also to my parents, for unconditional love and support, and for my "families" and friends, for sharing in a hug, a meal, my faith, or just my life. All in all, you have been a great blessing to me. And thank God.

> The fear of the Lord is the beginning of wisdom; all those who practice it have a good understanding. His praise endures forever!

Contents

List of Figures

List of Tables

List of Abbreviations

ADR Action Design Research
AE Accounting Entity
AER Accounting Entity Ranking
BIE Building, Intervention, and Evaluation
CART Classification and Regression Trees
CDA Continuous Double Auction
CSS Creativity Support Systems
DP Design Principle
DSR Design Science Research
DSS Decision Support System
EBS European Business School
EIX Economic Indicator Exchange
FSS Foresight Support System
FZI Forschungszentrum Informatik
GDSS Group Decision Support System
GWSS Group Wisdom Support System
HSX Hollywood Stock Exchange
ICT Information and Communication Technology
IT Information Technology
IS Information Systems
IQR Interquartile Range
IEM Iowa Electronic Markets
IISM Institute of Information Systems and Marketing
JFIM Judgmental Forecasting Improvement Model
KIT Karlsruhe Institute of Technology
LMSR Logarithmic Market Scoring Rule
MU Money Unit
NFL National Football League
OLS Ordinary Least Squares
PSM Political Stock Market
RTD Real-Time Delphi
sRTD social Real-Time Delphi
SD Standard Deviation
SSO Single Sign On
U.S. United States

Chapter 1

Introduction and Motivation

> I've read the last page of the Bible. It's all going to turn out all right.
>
> Billy Graham

1.1 Motivation

Prediction markets regularly experience short-term prominence in the public opinion and media when high-profile political events are approaching, such as the United States (U.S.) Presidential Elections or the German Federal Elections. They are considered a promising alternative and benchmark to polling. Though polling, prediction markets, and other forecasting methods (e.g., statistical models or Delphi studies) are regularly benchmarked against each other in public media and academic research, there is no consensus yet on "the one" most accurate method (e.g. Graefe, 2014; Graefe, 2017). In fact, since all methods have been steadily developed and improved over the last decades, the accuracy of one method in a previous election even seems to be negatively correlated with accuracy in an upcoming election (Graefe, 2014). Further, each method has its own strengths, weaknesses, and biases that depend on the individual context of a certain forecasting problem.

Steadily, good forecasts are becoming more important, not only in politics, but in many other areas, companies, and organizations (Durand, 2003; Blanc and Setzer, 2016). Thereby, accuracy is not the only measure and often other metrics (e.g. consensus) or more information (e.g. the reasoning behind a certain expectation or forecast) are required. Especially for decision makers that are potentially kept responsible for their decision outcomes, qualitative information on the background, relationships, and the reasoning behind certain opinions are important as well.

It has been long known that the combination of forecasts from different methods often yields superior or at least more robust accuracy (e.g. Graefe, 2014). In contrast, a more recent development is the combination or convergence of several methods itself. This approach does not only raise the forecast accuracy, but brings together individual advantages of each method in order to overcome certain challenges. One example are expectation surveys, combining polling and hidden markets[1] in order to make polling more robust against biased samples (George Mason University, 2015; Teschner and Weinhardt, 2012a). In this thesis, the concept of "Delphi-Markets"

[1] Hidden markets are prediction markets in which the market mechanism is concealed by an interface that only asks for expectations and associated parameters.

is introduced, representing such a novel combined approach that integrates prediction markets and Real-Time Delphis (RTDs)[2] – both feedback based approaches. The integration of these methods promises to address several shortcomings of each method, e.g., it allows enriching prediction market forecasts with qualitative feedback or addresses the expert selection process for Delphi studies. The approach was developed in the context of the DFG[3] supported project MInPuD[4] and made publicly available as the forecasting platform "FAZ.NET-Orakel" in cooperation with the online magazine of the Frankfurter Allgemeine Zeitung, FAZ.NET.

In the light of recent developments of the political environment and society, prediction markets are very likely to grow more important in the future. Therefore, it is relevant to keep improving this methodology by addressing some of its challenges. The current public opinion draws the picture of an increasingly politically polarized society (Yang et al., 2016). There is a link with emerging Information and Communication Technologies (ICTs) that are used more and more for political advertising (Greer and LaPointe, 2004; Lee et al., 2014). At the same time, certain political opinions are publicly stigmatized (Brug, Fennema, and Tillie, 2000), which is why a lack of "truth-revelation" can become (and already is) a real challenge in political research (Kennedy et al., 2017). In addition, traditional political opinion polls are expensive and chronically under-budgeted, which is only one reason for the insufficient quality and lack of necessary adjustments to usually distorted samples (Kennedy et al., 2017).

In this context, prediction markets have some very favorable properties. In contrast to polls that are usually used for election forecasting, prediction markets do not ask for the opinion or voting behavior of individuals, but for their expectations about the outcomes. This allows prediction markets to create forecasts that are more robust against distorted samples and also allows to create accurate prediction based on relatively few participants (Green, Armstrong, and Graefe, 2007). Further, prediction markets provide continuous forecasts that rapidly adapt to new information, while polls only reflect "snapshots" that may be quickly outdated (Graefe, Luckner, and Weinhardt, 2010). Other favorable properties are, e.g., anonymity or their potential incentive compatibility (Jurca and Faltings, 2008; Schlag, Tremewan, and Weele, 2015; Green, Armstrong, and Graefe, 2007). Therefore, there is also no doubt that prediction markets have enormous potential to generate and improve forecasts and the forecasting methodology of companies and other organizations (Buckley, 2016) - but there is also a need for improvement (Wolfers and Zitzewitz, 2006a). Several publications deal with deficiencies of prediction markets, errors and inaccuracies in their forecasts, manipulation, and systematic biases. Some weaknesses of prediction markets may be addressed by the integration of prediction markets with the RTD methodology, while RTDs have their own weaknesses that may be reduced by prediction markets. However, the detailed design of an integrated approach can have multiple forms, bearing individual potentials and challenges, which are outlined in this dissertation. One specific form is implemented in the FAZ.NET-Orakel and further refined. In the continual competition of forecasting methods and in face of recent developments regarding ICTs, society, and politics, it is of utmost importance to further investigate potentials and challenges of new approaches, such as the

[2] RTD is an approach for forecasting based on an expert panel. In several rounds the experts are faced to the same or related questions and to the aggregated opinion of the previous round. The approach aims to reach consensus among the experts. See Section 2.3 for further Details.

[3] Deutsche Forschungsgemeinschaft

[4] DFG sponsored project (WE 1436/12-1), "Methoden-Integration von Prognosemarkt und Delphi-Studie", (*engl.: Integration of the methods prediction market and Delphi study*)

suggested integration of prediction markets and RTDs, the Delphi-Markets.

In order to improve prediction markets, this dissertation highlights four potentials and challenges that were addressed in four distinct research projects and performed in the context of the prediction market FAZ.NET-Orakel or related to prediction markets in general. These four research projects shall improve the suggested integrated approach from a holistic perspective and are therefore derived from the four sources of error of the Judgmental Forecasting Improvement Model (JFIM), introduced and explained in detail in Section 2.1.3. The JFIM shows that forecasting errors are never only caused as a mere design problem of a certain method, but always in interaction of the method itself with a human factor (motivation and cognition). This also yields the structure of this thesis.

1.2 Structure of Dissertation

First, Chapter 2 explains relevant background information on prediction markets, RTDs, forecasting in general, and the research methods applied. Chapter 3 introduces and elaborates on the concept of Delphi-Markets with its challenges and potentials and describes the FAZ.NET-Orakel as one instantiation in detail. The following chapters improve the Delphi-Markets regarding the four errors identified in the JFIM. Chapter 4 highlights the potentials of the Delphi-Markets regarding the sampling error. The selection of experts for RTD studies is a current challenge to the RTD method and its rigor (Hasson and Keeney, 2011). It is evaluated with which accuracy the prediction market component can be used to select potentially knowledgeable participants to be invited as experts to the RTD survey. Chapter 5 highlights the problem of the non-response error. RTD studies regularly lack sufficient retention of the participants, which makes their results questionable and less generalizable (Cuhls, 2003). A social Real-Time Delphi (sRTD) approach is suggested and evaluated in order to raise retention in the FAZ.NET-Orakel artifact. Chapter 6 deals with the problem of cognitive biases. As all events have to be formulated as distinct shares, the partition dependence bias is especially problematic in prediction markets (Sonnemann et al., 2013). In this thesis it is shown that the prediction market design, as well as the selection of the participants (due to their properties, especially expertise) can have a significant influence on the occurrence of this bias. Chapter 7 improves Delphi-Markets and prediction markets in general regarding motivational biases, especially manipulation and fraud. Right from the launch of the FAZ.NET-Orakel in March 2017, manipulative and fraudulent attacks were recorded on the prediction market. This motivational bias is commonly known and recognized as one potential stumbling block for widespread practitioner adoption of prediction markets (Buckley and Doyle, 2017). This thesis improves the manipulation and fraud detection and prevention based on Action Design Research (ADR) with several design suggestions and principles. The crowd-based approach for manipulation and fraud detection suggests a solution for several problems in detection and handling of potentially manipulative or fraudulent cases. Each of the Chapters 4 to 7 is concluded with a discussion. However, Chapter 8 summarizes the results of each spotlight, discusses them on a meta level, and provides a short outlook into future work.

Chapter 2

Foundations

> In these democratic days, any investigation into the trustworthiness and peculiarities of popular judgments is of interest. The material about to be discussed refers to a small matter, but is much to the point.
>
> Galton (1907, p. 1)

2.1 Group- and Crowd-based Forecasting

2.1.1 A brief Introduction in Judgmental Forecasting

Judgment has always played an important role in forecasting (Lawrence et al., 2006). Though, in the last 30 years, it sometimes seems to have become commonplace among academics to warn against judgments in favor of statistical models, practitioners never shared this skepticism (Lawrence et al., 2006).

The errors of human judgment have been the subject of research of several famous researchers. Daniel Kahneman, an outstanding scientist and Nobel memorial price laureate, has dealt all his lifetime in his works on judgment under uncertainty with the heuristics of human information processing, and exemplifies his own devastating forecasting ability in several examples. During his national service in the Israeli Army, he was faced the task to predict cadets' success in the officer training. He and some colleagues observed the cadets in group tasks to assess their leadership skills, personality, communication capabilities, and so on in order to assess their success in an officers training. After a discussion on each cadet, he felt almost perfectly sure about each cadet's future, but his predictions finally turned out to be only a little better than blind guess (Kahneman, 2012). He became a victim of the "illusion of validity" bias, a heuristic that makes our brain jump to conclusions from only little evidence, not taking into account how big these jumps are. Heuristics and biases in human cognitive reasoning distort our judgment and trick our assessments and perceptions of probabilities and confidence (Tversky and Kahneman, 1974).

Another famous experiment by the renowned researcher in forecasting, Philip E. Tetlock, demonstrated in a large experimental study with a "[...] big group of experts – academics, pundits, and the like – [...] [and] thousands of predictions about the economy, stocks, elections, wars, and other issues of the day" (Tetlock and Gardner, 2015, p. 10) that the "experts" are no better than random guessing (Tetlock and Gardner, 2015). This experiment became famous with the "dart-throwing chimpanzee" comparison.

These and similar stories led to a general distrust in judgmental forecasting and a shift towards other, mostly statistical methods. With the rise of computers and better accessibility of data, statistical and quantitative models became more present in forecasting. However, these methods seem not to be fully capable to solve the challenges of forecasting. On the one hand several big failures become publicly known, such as the $400 Million experiment of Nike where wrong software forecasts led to massive inventory write-offs (Lawrence et al., 2006). On the other hand, they had limitations especially in the long-view as researchers and decision makers are often confronted with high uncertainty, new technologies, and new ideas frequently. In these situations analytic techniques and basic scenario methods often cannot be used, due to a lack of specific information or technical and historical data (Linstone and Turoff, 2002c). However, reliable forecasting and assessment of future developments have always been key success factors of governments, companies, and organizations and are becoming more and more strategic resources (Durand, 2003).

In the light of the challenges mentioned above (and some more), many researchers tried to improve forecasting – judgmental, statistical, and quantitative. In short- and mid-term forecasting, and where historical data is present, combined approaches find vast application (Lawrence et al., 2006). For instance, Decision Support Systems (DSSs) access historical data and other data sources and generate a forecast that is then often revised and adapted in a second loop by a human forecaster (Lawrence et al., 2006) – or the other way around (Knöll and Simko, 2017). In mid- and long-term forecasting or in forecasting where no historical data is available, methods were developed to help forecasters in considering all relevant information and to reassess and adapt their evaluations, simultaneously overcoming systematic biases. Those are, e.g., the Delphi method (e.g. Linstone and Turoff, 2002c), prediction markets (e.g. Luckner et al., 2012), or other methods[1] (e.g. Atanasov et al., 2016; Rothschild and Wolfers, 2013). All these methods are judgmental forecasting methods. Though there is no universal definition, one may consider all methods as judgmental forecasting methods that are mainly based on subjective information of individuals (*Judgmental Forecasting* 2009). Sometimes, attempts for definitions use phrases like "expertise based on experience" or "relying on subjective information" (e.g. Armstrong and Green, 2018), however, most definitions also mention that this subjective forecast may also rely on quantitative data that is processed by the individual. Therefore, this thesis applies the following working definition on judgmental forecasting:

> *Forecasting methods in which a human individual or group/crowd processes underlying and potentially dispersed information in order to generate a forecast or at least a part of a forecast are considered as methods of "judgmental forecasting".*

Judgmental forecasting methods also include intention surveys or intention polls that do not collect individual expectations, but individual plans or behavior (Armstrong, 1985). However, the focus of this work is on those methods that collect individual expectations: Expectation based methods.

However, there exists a variety of expectation based methods in judgmental forecasting and different contexts favor different methods, as they are, e.g., more prone to certain errors. Their overall forecasting error is driven by several "sources" and

[1] E.g., expectation polls or peer prediction schemes.

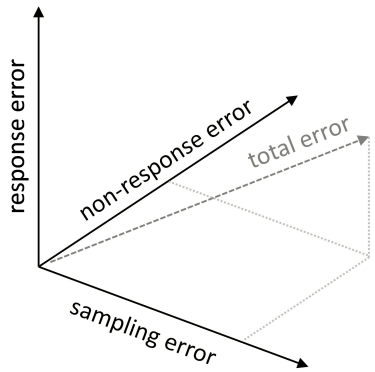


FIGURE 2.1: Decomposition of the total forecasting error in opinion based forecasting (a three-dimensional graph). Adapted from Armstrong (1985, p. 83).

some methods deal better with certain errors than others. Armstrong (1985) decomposes the overall forecasting error in judgmental forecasting into a three-dimensional error model which is illustrated in Figure 2.1. Besides an error in the expectation of each forecaster itself (individual variance of estimation), the three-dimensional model distinguishes between (1) response error, (2) non-response error, and (3) sampling error. These components are arranged as three dimensions in a room to illustrate that they are largely independent of each other. The total error can then be interpreted as the vector sum of each component. Armstrong initially defined the model in 1985 for "opinion-based forecasting", which would now, more than twenty-five years later, correspond to what is currently referred to as "expectation-based forecasting" and is, therefore, regarded as valid for judgmental forecasting in general.

In the following, the three components are described more in detail.

The **response error** describes a wide field of errors that are caused by individual biases or decisions of individuals. According to Curtis and Wood (2004) and Skinner (2009) such biases can be divided in (i) cognitive biases and (ii) motivational biases.

Cognitive biases are the result of decisions based on simple heuristics or "rules of thumb" (Evans, 2012; Tversky and Kahneman, 1974) and also include those biases and heuristics described by Kahneman (2012). Probably the best known cognitive biases are the anchoring and adjustment heuristic (Tversky and Kahneman, 1974) or the status quo bias (Kahneman, Knetsch, and Thaler, 1991). A prominent representative in economics is the endowment effect or the quasi-endowment effect (though less prominent in judgmental forecasting) (Kahneman, Knetsch, and Thaler, 1991; Gimpel, 2007). Overall, there are more than 100 cognitive biases (Benson, 2016). According to Müller-Trede et al. (2018) one big advantage to utilize crowds for forecasting is the basic idea that individual errors of the forecasters are averaged out. And there exists large evidence confirming this for many biases. However, several studies, among others in the context of prediction markets, showed that expectation based forecasting is not excepted from these errors. Currently, many studies address cognitive biases in decision-making (e.g. Jung and Dorner, 2018) and forecasting,

both in RTD studies or traditional Delphi studies[2] (e.g. Winkler and Moser, 2016) and prediction markets (e.g. Sonnemann et al., 2013).

Motivational biases are errors that occur if individual participants have some interest in changing, distorting, or biasing the results away from their true values(Curtis and Wood, 2004) or other incompatible interests that distort the prediction. Many of these phenomena are better known under the term manipulation (and fraud) in prediction tasks. They may, however, also be caused by fear of stigmatization (e.g., a participant does not reveal his true belief, as it fears that its opinion is socially not accepted) or untruthful revelation because of other reasons (e.g., if anonymity is not guaranteed or not perceived) (e.g. Kennedy et al., 2017).

The **sampling error** describes the problems when the sample does not adequately represent the population. Although there is consent that expectation-based forecasting is much less prone to sampling errors (Rothschild and Wolfers, 2013; Winkler and Moser, 2016), there are several studies that demonstrated the judgment bias in Political Stock Markets (PSMs)[3] (Forsythe, Rietz, and Ross, 1999; Berlemann and Schmidt, 2001; Graefe, 2014; Kranz et al., 2014). In prediction markets, as an instance of expectation- and crowd-based methods, the sampling error is existent and has a measurable, though small impact on the forecast accuracy. However, in expert-based judgmental methods, such as the Delphi method, the selection of the experts has been identified as a key determinant for the success, the validity of the results (accuracy), and the broadness of the discussed opinions (Welty, 1972; Hill and Fowles, 1975; Green, Armstrong, and Graefe, 2007).

The **non-response error** describes all problems that arise from non-respondent subjects. The non-response error often also leads to a sampling errors. A sample may adequately represent the field and information available; however, if not all participants answer, a skewed sample may be the result (Armstrong, 1985). This phenomenon is also referred to as self-selection bias (Heckman, 1990). Basically, this problem exists in all group- and crowd-based forecasting methods and occurs at the beginning of a forecasting process as well as during the process. During the process, such cases are called drop-outs: Subjects that initially participated in the process, but do not answer or respond to subsequent parts of the process. In the Delphi method, drop-outs are a persistent problem and still a challenging factor for the success of studies (Mullen, 2003; Reid, 1988; Walker and Selfe, 1996).

It is to note that these types of error are not independent of each other. Self-selection, as a non-response error, ultimately leads to a sampling error. Sampling errors may be caused by cognitive biases, such as the judgment bias. In many cases it is not possible to analyze solely one type of error. Therefore, the evaluation of forecasting methods should always consider different perspectives and should be performed in a holistic approach.

2.1.2 Judgmental Forecasting Methods

As already mentioned in Section 2.1.1, there are various validated and applied forecasting methods. Each method has its own strengths and weaknesses in different contexts. Armstrong and Green (2018) identified 17 valid forecasting methods that

[2]The Delphi method is likewise the already introduced RTD method an approach for forecasting based on an expert panel relying on the key properties anonymity, iteration, controlled feedback, and statistical group response.

[3]PSMs are prediction markets with a focus on the outcome of election and political decisions.

are reported and applied in literature and practice. Among these are nine judgmental methods and five quantitative methods. Two further meta-methods are constructed as combinations of single or different methods. The judgmental methods are the following (Armstrong and Green, 2018, p. 4):

- Prediction markets[4]
- Multiplicative decomposition[5]
- Intentions surveys[6]
- Expectations surveys
- Expert surveys (Delphi, etc.)[7]
- Simulated interaction[8]
- Structured analogies[9]
- Experimentation
- Expert systems[10]

Armstrong and Green (2018) argue that often forecasting methods are only applied to specific problems for the reason that the conducting researcher is familiar with the method. E.g., many students become familiar with polls and the related methodology during their time at university. If confronted with a forecasting task in their later life, polling seems a promising option, as they know and understand the methodology and the necessary steps. Nevertheless, an expert based approach or a prediction market may have been more fitting to the context. Due to a lack of knowledge and effort in the consideration of advantages and disadvantages, the student will still use polls. For this reason, Armstrong and Green (2018) created a check list that shall be used to compare all forecasting methods and to find the best for each specific forecasting task. These methods can be understood as "archetypes" of judgmental forecasting.

Finally, in many cases different models fit a specific problem and even if the assumable best model is found, due to very high variances in forecasting accuracy, another model might have performed better. Graefe et al. (2015) demonstrated on historical data from the U.S. Presidential Elections 1976 to 2012 that the high accuracy of a method in a preceding election was negatively correlated with high accuracy of the same method in the upcoming election. This basically means that a method that performed good on a certain forecasting task in the past may still perform bad in the same task in the future (or at least worse than other methods[11]). To select the best method for a certain context is, therefore, arguably hard and sometimes even subject to "luck". In many use-cases, e.g., elections, a variety of methods would be applicable and potentially successful.

[4] Explained in detail in Section 2.2.

[5] The forecasting problem is represented as a product of smaller forecasting problems.

[6] E.g., election polls.

[7] Explained in detail in Section 2.3.

[8] A form of role-playing that is used to forecast decisions by people, e.g., in conflict situations. Naïve subjects are introduced to specific roles and instructed to engage in realistic interactions until a decision is reached (Armstrong and Green, 2018).

[9] Experts are asked to suggest situations that were similar to the target situation, rate their similarity as well as map outcomes in the suggested situation to outcomes in the target simulation. Afterwards, a forecast is calculated based on the provided estimations on similarity and outcomes (Armstrong and Green, 2018).

[10] Experts describe their process of making forecasts step-by-step, such that it can be implemented as software (Armstrong and Green, 2018).

[11] It has to be noted, that the ongoing development and improvement of the methods (also by a raise in computing power and available data) may effect these high variances in forecasting accuracy.

2.1.3 The Judgmental Forecasting Improvement Model (JFIM)

The three-dimensional error model from Armstrong (1985) identifies sources of error for group- and crowd-based judgmental forecasting approaches. However, the impact of certain sources of errors often cannot be mapped directly to the outcome accuracy. If a sample error occurs in an expectation based judgmental forecasting method, you cannot simply calculate how much better the accuracy would have been if the sample was not distorted. Even an experimental setup would not finally solve this problem, unless it would be carried out over a very long period, as variance in forecasting is too large. This was demonstrated by Graefe et al. (2015) on historical election forecasts. The same argumentation is valid for motivational factors: Though it definitely leads to an error in the forecast, this often cannot be quantified properly. Therefore, an investigation on the advantages of a new forecasting approach, such as the integrated approach of the Delphi-Markets that are introduced in this thesis, has to be evaluated on other quality measures that better fit each individual improvement.

For this reason, this dissertation develops and applies the Judgmental Forecasting Improvement Model (JFIM) as a meta model to identify starting points for improvement. The JFIM maps the core aspects of a judgmental forecasting method by Lyon and Pacuit (2013) (adjustable by the researcher) to the external factors of the participants and to the four types of error discussed above. Lyon and Pacuit (2013) identified six "core aspects" of wisdom of the crowd in the context of individual judgment, judgment aggregation and collective judgment. These six aspects are capable to describe a crowd-based (judgmental) forecasting application or platform from a design perspective:

- (1) The output.
- (2) The recruitment.
- (3) The inputs.
- (4) The elicitation method.
- (5) The aggregation method.
- (6) The standard of evaluation.

In order to improve judgmental forecasting methods, it is possible to adapt and modify these design elements or decisions.

Figure 2.2 reassembles the core aspects by Lyon and Pacuit (2013) graphically, but from a Information Systems (IS) design perspective:

- The elicitation (4) and aggregation (5) assemble the forecasting mechanism (method) and are therefore illustrated as an underlying platform (box).
- Illustrated as a black circle in the center, the output (1) is defined as the target of the underlying forecasting project and fully based (dependent) on the mechanism.
- The standard of evaluation (6) is a added as vertical box to illustrate its independence of the forecasting mechanism. Its shape (box) indicates its indirect influence on the outcome, analogically to the elicitation and aggregation element.
- The recruitment (2) and inputs (3) are the interfaces of the crowd's estimation to the forecasting method and finally to the output. Their shape indicates their direct influence on the output. As they are dependent of the underlying forecasting mechanism they intersect with the boxes of elicitation and aggregation.

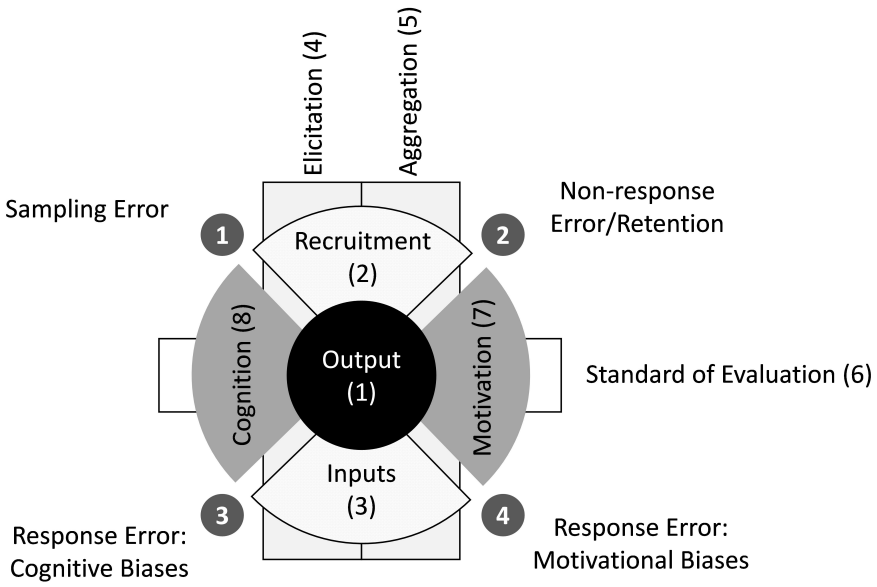


FIGURE 2.2: Judgmental Forecasting Improvement Model.

However, in contrast to the elicitation (4) and aggregation (5), they are much more question specific.

However, neither the recruitment nor the inputs are completely unbiased. The motivation (7) of the crowd and its individuals, as well as their cognition processes (8) interfere with the forming and submission of the expectations and also have an impact on the output, the final forecast (Kloker and Kranz, 2017; Watts, 2015; Evans, 2012). Analogical to the recruitment and inputs, the motivation and cognition have a direct influence on the output. However, cognition is in part moderated by the elicitation, for which reason the two shapes intersect (Does the elicitation method provide framing, anchors, or other heuristic cues?). The motivation is also influenced by the aggregation (Does an individual estimation really matter? Or is it concealed by the aggregation?). Therefore, the two shapes also intersect.

This way of assembling the core design elements and crowds properties of a judgmental forecasting method provides some further value: When this new model is integrated with the three-dimensional error model of Armstrong (1985), the error components can be located in the touch points of the four quadrants (see Figure 2.2) as they are often the result of the interaction between design elements and the crowds' properties. The JFIM is therefore capable to give at least a first reference point, where the design has to be adapted in order to reduce certain biases or errors.

- To reduce the sampling error one have to consider if the recruitment processes leads to distorted samples. For opinion based judgmental forecasting methods, such as polls, this would mean to select a sample in a representative manner. For expert based approaches of judgmental forecasting or prediction markets, it is necessary to select the participants according to their cognition (e.g. expertise) in order to have an optimal sample.
- To reduce the non-response error one may intend to adapt the recruitment processes and or the motivation of the participants. This is possible by providing

external incentives, but also by, e.g., increasing intrinsic motivation using social cues.

- To reduce the cognitive biases (response error) one may consider how the cognition and the inputs are influenced by the elicitation mechanism and the way the inputs have to be provided.
- To reduce motivational biases (response error) it is necessary to monitor the inputs provided by the participants to detect fraud or manipulation. But it is also necessary to enrich the forecasting method (or application) with preventive design elements that decreases harmful motivation.

In order to improve the integrated approach introduced in this thesis, this model will be applied and all four touch points will be addressed within an individual research project.

2.2 Prediction Markets

2.2.1 A brief History of Prediction Markets

The "wisdom of crowds" is the basic principle underlying all prediction markets. The term was coined by Galton and his experiment during the show of an annual fat stock and poultry exhibition in 1906 (latter referred to as the "Galton Experiment") (Galton, 1907; Shrier et al., 2016). For 6 pence persons could buy stamped and numbered cards and submit their estimate on the weight of a specified ox. The average competitor was not likely to be an expert in oxen, not more than a voter on the political issues that he casts his vote on. However, to the surprise of Galton, the "vox populi" (voice of the people) was astonishingly accurate. The median was 0.8% away from the true value, the average only little more (0.85%) (Shrier et al., 2016). Galton (1907, p. 5) concludes that "[...] [this] result is [...] more creditable to the trust-worthiness of a democratic judgment than might have been expected.".

The "wisdom of crowds" is explained statistically with an additive error model (Shrier et al., 2016). If an individual forecast Y_i is a function of the true value x and an individual error $\epsilon_i(x)$ and this individual error is unbiased and independent ($E_x[\epsilon_i(x)] = 0 \ \forall i$), the average forecast should be very close to the true value x if the sample is large enough. However, by simply averaging all estimations, the accuracy is only given if the individual forecasts Y_i reflect the true estimation of each individual forecaster. Markets are an aggregation mechanisms that usually does not have the problem of untruthful revelation of information, as they yield expected financial losses.

Another famous example in the history of prediction markets occurred in 1984 in the U.S.. An examination of the relationship between the weather and the orange futures revealed that the price of the futures at market closing predicted the errors in the weather forecasts of the minimum temperature in the later evening (Shrier et al., 2016). Orange trees cannot deal with temperatures below zero degrees Celsius for more than a few hours. Orange farmers, therefore, observe the weather very carefully. Based on their observations and experience they often can make better estimations on the local weather than any centralized weather models. In order to make profit, they trade on these orange futures as well and their knowledge becomes reflected in the price of the future. The example of the orange futures is therefore a good example that the crowd is capable to collect and hold a huge amount of public and private, and centralized and dispersed information (Shrier et al., 2016). Markets are capable to give the crowd an incentive for revealing this information.

In the founding article "The use of knowledge in society", Hayek described this property of markets: "Fundamentally, in a system where the knowledge of the relevant facts is dispersed among many people, prices can act to cordinate [!sic] the separate actions of different people in the same way as subjective values help the individual to cordinate [!sic] the parts of his plan" (Hayek, 1945, p. 526). Markets provide people with incentives to use their information, while observing the feedback on the opinion of the other participants. Valid information helps participants to detect over- or under-valuations in the market and allow the participant to realize profits by trading. The trading activity, however, leads to changes in the current market price and therefore reflects the new information to all market participants and observers. With the "efficient markets hypothesis", Fama (1970) formulated the fact that all available and relevant information is reflected in the market price. There are different levels of strictness, but in the strong form of the efficient markets hypothesis, it is said that the price would be unaffected by any disclosure of information that was previously known to a person (Malkiel, 2003; Sewell, 2011). It is not to assume that (all) prediction markets are strong efficient markets, but according to Snowberg, Wolfers, and Zitzewitz (2005, p. 370) prediction markets "[...] seem to satisfy at least the weak form of the efficient markets hypothesis". Basically, this means that the revelation of private information may affect the market price, but there are no profit opportunities from applying strategies that simply rely on past prices (price follows a random walk)[12]. Based on this principle, prediction markets function basically according to the following process, illustrated in Figure 2.3: (1) The crowd shares some common information, as well as every individual collects and has some private information. According to this information each individual forms its own expectation. (2) Based on this expectation, each individual can calculate how much it values one share in the prediction market. If they perceive the price as to low, they buy shares, as this results in an expected profit. Otherwise, if they perceive the price as to high, they sell. (3) This trading results in a price that continuously adapts to changing information according to the beliefs of all individuals. (4) It also serves as a public information to the crowd on the crowds' belief.

Prediction markets became famous with the introduction of the Iowa Electronic Markets (IEM) (Berg, Forsythe, and Rietz, 1997). Since 1988 it allows students to trade on the outcomes of diverse events in politics and finance. Especially prominent are the forecasts for the U.S. Presidential Elections, as they outperformed large-scale polling organizations (Wolfers and Zitzewitz, 2004). Over time, however, their application spread above many other contexts.

2.2.2 Contexts in which prediction markets were used

Nowadays, prediction markets are used in a variety of fields (Luckner et al., 2012). For instance, Table 2.1 gives a brief overview on prediction markets and their contexts. Many prediction markets cannot be sorted in one single category. Therefore, Table 2.1 should just give an impression and should not be regarded a complete and exclusive list. Prediction markets experienced their greatest visibility in the area of politics. However, in sports and alongside sports betting they are also regularly applied. Application in companies and organizations (e.g., for DSS, Foresight Support System (FSS), Idea/Innovation/Project Management) is currently becoming more and more popular (Prokesch, Gracht, and Wohlenberg, 2015; Buckley, 2016;

[12] A third variant of the efficient market hypothesis is the semi-strong from efficiency. Here prices adjust rapidly to publicly available information. These adjustments are unbiased, so no excess returns are possible by trading on this information.

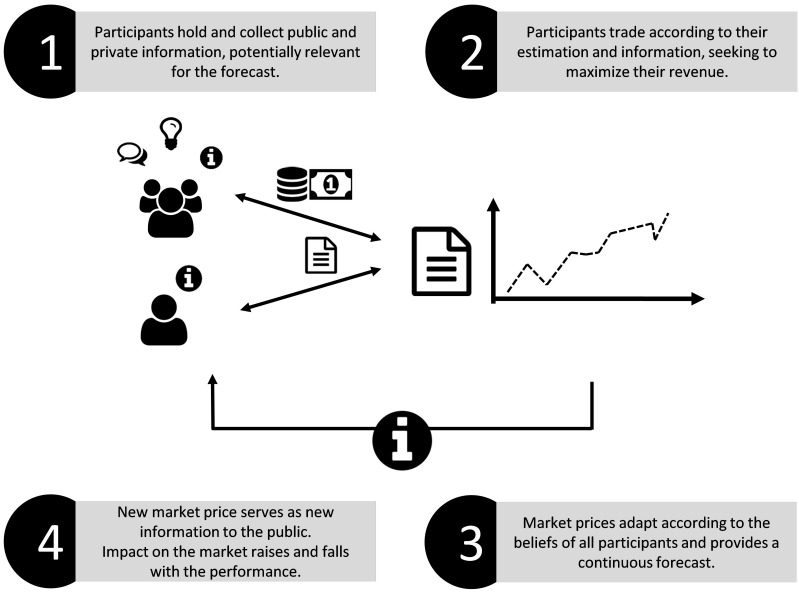


FIGURE 2.3: Basic process of a prediction market. Participants collect dispersed information and trade according to their expectations on the market. The market price serves as new information to the traders.

Soukhoroukova, Spann, and Skiera, 2012; Remidez Jr and Joslin, 2007). Other applications, such as science, are less prominent. Prokesch, Gracht, and Wohlenberg (2015) argues strongly for their application for economic indicators and Teschner, Stathel, and Weinhardt (2011) found that prediction markets perform at least competitive with Bloomberg forecasts. Accordingly, the Handelsblatt and the Frankfurter Allgemeine Zeitung (two large German news magazines) featured and still feature prediction markets for economic indicators within the context of their online news magazines.

2.2.3 Interpreting prices as forecasts

As to Figure 2.3, in prediction markets forecasting issues are modeled in contracts, which define the rule how the market will pay out the traders regarding the realization of an event. One reason for the large applicability of prediction markets is the fact that contracts can be modeled in a way that the prices can be directly interpreted as probabilities with respect to the event to be predicted (Wolfers and Zitzewitz, 2006b). This makes it easy for participants to translate their expectations into trading. There are, however, different types of contracts. These are (1) winner-take-all contracts, (2) index contracts, and (3) spread contracts (Wolfers and Zitzewitz, 2004).

- **Winner-take-all contracts** pay out the maximum amount (often 100 or 1 Money Unit (MU)) if an event happens, otherwise nothing. "Will candidate XY win the election?" is such an example. The market price of a contract $c(y)$ for an event y can then be directly interpreted as the probability $p(y)$ that event y occurs.
- **Index contracts** pay according to a certain rule based on the realization of an event. "How much percent will candidate XY achieve?" is such an example. A rule would, e.g., define that the contract pays 1 MU for each percent

Area	**Examples**	**References**
PSM	IEM (rm), Wahl$treet (c), PESM Wahlboerse (rm), Wahlfieber, Pollyvote Prediction Market, Economic Indicator Exchange (EIX) (c), Predictit (rm)	Luckner, Kratzer, and Weinhardt (2005), Graefe (2017), Franke, Geyer-Schulz, and Hoser (2005), and Franke, Geyer-Schulz, and Hoser (2006)
Sports	STOCCER (c), TradeSports (c, rm), Smarkets (rm), Betfair (rm)	O'Connor and Zhou (2008)
Economic Indicators	EIX (c), Kurspiloten (c)	Teschner, Stathel, and Weinhardt (2011)
Science	DAGGRE (c), SciCast (c),	Dreber et al. (2015), Laskey, Hanson, and Twardy (2015), and Powell et al. (2013)
Entertainment	Hollywood Stock Exchange (HSX), Fame Project (rm)	Pennock et al. (2001)
All of the above	Intrade (c, rm), Hypermind (rm), FAZ.NET-Orakel, Predictious (rm, bc), Augur (rm, bc), Metaculus, NewsFutures (c)	Bruneel et al. (2018) and Peterson et al. (2018)
DSS, FSS, Idea/-Innovation/Project Management	Google, HP, Deutsche Telekom, Yahoo, General Electrics	Cowgill, Wolfers, and Zitzewitz (2009), Prokesch, Gracht, and Wohlenberg (2015), Buckley (2016), Van Bruggen et al. (2010), Rohrbeck, Thom, and Arnold (2015), Keller and Gracht (2014), Soukhoroukova, Spann, and Skiera (2012), and Spears and LaComb (2009)
Others	Climate Change, Swine Flu Pandemic, Research Evaluations, Education, ...	Munafo et al. (2015) and Buckley and Doyle (2015)

TABLE 2.1: Selected popular prediction markets and their area of application. (c = closed, rm = real money, bc = blockchain-based)

of vote-share. The market price of a contract $c(y)$ for an event y can then be interpreted as the mean value of the realization of this event $E[y]$.

- **Spread contracts** pay according to a certain rule based on a property of the expectations on the realization of an event. Usually this is, that a contract doubles its price if the realization of an event y is larger (smaller) than a certain threshold $y*$ and pays 0 if it is below or equal (above or equal) this threshold. This contract design allows to elicit the median value of the expectations from the participants (Wolfers and Zitzewitz, 2004).

2.2.4 Advantages of Prediction Markets

Prediction markets feature several further beneficial properties that are the reason for their wide application: (1) fast reaction to new information (Graefe, Luckner, and Weinhardt, 2010), (2) continuous forecasts (Graefe, Luckner, and Weinhardt, 2010), (3) immediate feedback, (4) easy implementing of incentives (partly incentive compatibility) (Jurca and Faltings, 2008), (5) implicit long-term weighting (Arrow et al., 2008), (6) gamification is possible (Buckley and Doyle, 2017), and (7) the collection of expectations (Rothschild and Wolfers, 2013). A detailed investigation of these and further advantages, as well as the disadvantages, will be given in Section 3.1.

2.2.5 Ongoing Discussions and Current Developments

As many other forecasting methods, prediction markets are continuously evaluated and developed. For instance, Kranz (2015) adapted the market-engineering framework (Weinhardt, Holtmann, and Neumann, 2003) and introduced "continuous market-engineering" as a process to continuously enhance and adapt markets to changing environments. Wolfers and Zitzewitz formulated in 2006 the "Five Open Questions About Prediction Markets", which then where the (1) attraction of uninformed traders, (2) limited contractibility, (3) manipulation, (4) calibration for small probabilities, and (5) separating correlation from causation (Wolfers and Zitzewitz, 2006a). This is a not full list of issues that have driven, and still do, current discussions and research on prediction market yet, of which a few shall be highlighted in the following paragraphs.

Accuracy

The probably never ending "ongoing discussion" is regarding the accuracy of prediction markets. In the field of politics, there is almost common sense that prediction markets produce competitive results (compared to polls), which was demonstrated by various studies (Wolfers and Zitzewitz, 2004; Berlemann and Schmidt, 2001; Rothschild, 2009). First euphoric studies that attributed prediction markets with a superior forecasting capability (e.g. Berg, Nelson, and Rietz, 2008) could not be positively reevaluated on following elections. In recent elections, prediction markets rather show an alternating success. While in the 2016 U.S. Presidential Elections, prediction markets did not "succeed" (Graefe, 2017), they were among the most accurate methods in the 2017 German Federal Election. The accuracy of prediction markets is usually bench-marked against polls and other forecasting methods. And in fact, since all methods have been steadily developed and improved over the last decades, the accuracy of one method in a previous election seems even to be negatively correlated with accuracy in an upcoming election (Graefe, 2014). Even between prediction markets with different market mechanisms there exist significant

differences regarding accuracy. Klingert and Meyer (2018) found in simulations that some mechanisms (e.g., Continuous Double Auction (CDA)) result in better accuracy levels than others. This was, however, also strongly dependent on the actors and the environment.

There is an ongoing discussion, if prediction markets do not aggregate private information but only reflect the results of the latest polls. This, however, would not suggest the application of prediction markets in a context where no public polls are available or when polls are not reliable[13]. Erikson and Wlezien (2012) conclude on this discussion that the truth lays somehow in between. According to the findings of Erikson and Wlezien (2012) prediction markets "worked remarkably well" before polls are available. However, as soon as public opinion polls are available the election markets seem to follow the polls. In the case of the Brexit, Fry and Brint (2017) could not find supporting evidence for this statement, in fact they appeared remarkably unresponsive. One may argue for both sides, but the hypothesis that markets only aggregate poll information, which would imply that they only work when polls are available, has to be rejected. Prediction markets put the question in a different way than the usual voter intention polls. Prediction markets ask for the expectation of each participant on the outcome of the event. Rothschild and Wolfers (2013) are comparing intention polls and expectation polls and conclude that expectation polls yield better results than intention polls. They describe the effect that the answers one receive in expectation polls are "[...] about as informative as if they were themselves based on a [...] [intention] poll of approximately twenty friends, family, and coworkers" (Rothschild and Wolfers, 2013, p. 41). Same should be valid in prediction markets. While in intention polls, every participant has to reveal his voting preference, the trader does not have to reveal her preference and can keep "anonymity". The superiority of expectation polls was also robust when biased samples were used (e.g. in the context of U.S. Presidential Elections, only democrats or only republicans) (Rothschild and Wolfers, 2013). Graefe (2014) is also comparing opinion surveys to expectation surveys as well as to prediction markets, expert forecasts, and quantitative methods. The dominance of the expectation surveys over the opinion surveys was also a robust finding in his studies. A weak dominance of the expectation surveys over prediction markets was demonstrated by Miller et al. (2012) (2008 U.S. Presidential Election, 19.000 respondents vs. Intrade prediction market). Graefe (2014) attributes these findings to a biased sample of traders in the IEM (white, male, well-educated, middle- and upper-income, in majority conservative voters) as well as other biases. However, Graefe (2014) concludes that expectation surveys and prediction markets are, regarding their accuracy, at least competitive models. Both findings support the hypothesis that prediction markets do not only aggregate public opinion polls but also the opinion, the "common sense", of their private environment and relationships (peer group).

Graefe (2011) compared the accuracy of prediction markets in a meta analysis of previous studies in the context of companies. Graefe (2011) showed that prediction markets performed better (+6% accuracy) than econometric models, but only competitive to other methods of combined judgment. However, Graefe (2011) also concludes that the heterogeneity of application areas and the lack of more studies makes it very difficult to evaluate this matter on a more resilient basis. Some studies were distinctly more accurate (+28%), others clearly worse (-29%) compared to the

[13] In domains other than politics this would be bookmaker odds, expert forecasts, etc. (Spann and Skiera, 2009; Straub, Teubner, and Weinhardt, 2016)

benchmarks. Regarding economic indicators, Teschner (2011) demonstrated competitive accuracy compared to the Bloomberg publications, sometimes even weeks before. Spann and Skiera (2009) assessed the accuracy of prediction markets against betting odds and tipsters in the area of sports. While prediction markets and betting odds performed competitive, tipsters were significantly worse. Atanasov et al. (2016) assessed prediction markets in the context of geopolitical events. While prediction markets performed continuously better than simple averages of individual forecasts, other aggregation mechanisms performed at least competitive or better.

It is to conclude that prediction markets are no panacea for forecasting problems, but are in many cases a competitive forecasting method that provides many other beneficial properties besides high accuracy. In addition, in areas besides politics, sports, and economics, where the outcomes can be observed objectively, evaluation may also be subject to manipulation or other strategic influences (Munafo et al., 2015; Prokesch, Gracht, and Wohlenberg, 2015). All in all, assessment of prediction market accuracy remains, often due to a lack of benchmarks, difficult in many areas, such as, e.g., idea management and evaluation.

Real Money vs. Play Money and other Incentives

Proper incentives are a key element and challenge in prediction markets. In prediction markets run on real money, these proper incentives are usually given: Good predictions causes gains, bad predictions causes losses and every participant is, therefore, given an incentive to trade according to his true beliefs (Blume, 2012). However, in many countries real money prediction markets are considered as gambling and would therefore require a gambling license, which is usually not the case for prediction market providers. Play money prediction markets are used instead. In play money prediction markets, however, bad predictions do not cause real (financial) losses. And good predictions also cause no real (financial) gains. As proper incentives are not given any longer, new incentive schemes are used, such as prizes for the top ranked traders or lotteries among all participants (Luckner, 2006). There is, however, the question if these incentive schemes, which do not necessarily pay off in a linear relationship to the performance and usually do not punish bad predictions, still perform well. Servan-Schreiber et al. (2004) addressed this question and compared results from a real money prediction market (TradeSports) to a play money prediction market (Newsfutures) in the context of the National Football League (NFL). Their findings indicated that there is no positive nor negative effect on the accuracy. Servan-Schreiber et al. (2004), however, noticed that knowledge and motivation are two essential factors for the accuracy and "[...] traders on both websites are obviously motivated and, at least in general, knowledgeable about the issues being traded" Servan-Schreiber et al. (2004, p. 250). In a later study, Servan-Schreiber (2017) demonstrated that there was also no difference in the field of politics. Gruca, Berg, and Cipriano (2008) also found no difference in the field of entertainment and Slamka, Soukhoroukova, and Spann (2008) found no difference in diverse areas at least regarding the reaction to new information. Rosenbloom and Notz (2006) and Diemer and Poblete (2010), however, found a difference in favor of real money prediction markets in non-sports events and "diverse events". Servan-Schreiber et al. (2004) argues that money is only one way to motivate knowledgeable participants. Other methods may be community, bragging, rights, or prizes. The way of motivation may, however, have an "[...] impact as well on the kind of person that registers to trade" and while real money may better motivate information discovery, play

money "[...] may yield more efficient information aggregation" (Servan-Schreiber et al., 2004, p. 250).

Luckner (2006) and Luckner and Weinhardt (2007) examine how a proper incentive scheme in play money prediction markets has to be designed. Interestingly, they found that performance-compatible payment schemes, where participants could "lose" money, performed worse than a fixed payment scheme or rank-order tournaments. Both studies suggest the rank-order tournament as the dominant payment scheme (regarding accuracy).

However, in rank-order tournaments a special problem occurs: All one-shot winner-take-all contracts require a counterpart contract that makes the overall market more complex for the participants. Assume a contract for an event Y: "Will candidate XY win the election?". Now assume that the average expectation of all traders is 80%: $E(y) = 0.8$; which is the most likely value. Let the estimation of the traders follow a normal distribution around 0.8: $E_i(x) = \Phi(x)$. A first trader k has a very good estimate on the most likely values and trades shares around 0.8. A second trader j has an estimate that is too low and will sell all shares until 0.3. A third trader l has an estimate that is too high and will buy all shares until 0.9. In this setting, though trader k has the best estimate, either trader j or l will receive the highest payout, as the contract will either pay 1 or 0. In one-shot markets it is therefore always better to take the risk and bet "double or quits". To solve this problem, it is necessary to introduce a counterpart contract $\neg y$: "Will candidate XY *not* win the election?". Trader k will trade shares around 0.2. Trader j will buy all shares until 0.7. Trader l will sell all shares until 0.1. However, both contracts combined, trader k will now have the highest payoff. The counterpart contract $\neg y$ solves the problem, however, leads to the fact that every trader has to trade his preference twice. In an attenuated form, the same problem exists in one-shot index and spread contracts as well.

Market Scoring Rules and Combinatorial Prediction Markets

Traditional prediction markets face diverse problems, such as low liquidity (Li and Vaughan, 2013) or the difficult modeling of dependent events (Powell et al., 2013). Market makers are one solution to this problem. Market makers offer for each security a price (or two prices) to which they are willing to accept buy and sell orders. Therefore, each participant is instantly able to trade. These prices are usually calculated based on market scoring rules. The most popular is the Logarithmic Market Scoring Rule (LMSR) by Hanson (2002), also called: "Hanson Market Maker". Proper scoring rules are incentive compatible, which means that only the true report of an agent maximizes its expected profits (Hanson, 2002)[14,15]. Though many positive properties, market scoring rules still have open questions. Many, including the LMSR, are dependent on a liquidity parameter that basically moderates the

[14] A market is incentive compatible if traders cannot obtain a higher profit through not revealing their true preferences, fraudulent actions, or if participants avoid participation (also, as they fear manipulation). According to the no-trade theorem, no rational trader should trade in a CDA market. However, the presence of noisy traders in a market makes the CDA a positive sum game for the rational traders and therefore "individual rational"(Kyle, 2016). In contrast, prediction markets based on scoring rules are only incentive compatible in the short run. Intelligent traders understand the influences of their trading on other participants and can realize more profits by first misleading the market participants by trading into the false direction and later trade according to their true preferences (Chen and Pennock, 2010).

[15] If prediction markets and or market scoring rules are really incentive compatible, especially in the long term, is an ongoing discussion. Ban (2018) argues that the LMSR is not incentive compatible but suggests three further adaptions which fulfill all requirements.

effect of a single trader on the market price. Setting this parameter, which has a significant influence on the forecasting accuracy, has long been perceived rather as art than science (Pennock, 2010). Karimi and Dimitrov (2018) only recently suggested an approach to calculate the optimal liquidity parameter. Due to the lack of real world data based evidence, it cannot be concluded yet, if this approach works for all prediction markets.

The LMSR has one further unique advantage that is especially important for combinatorial prediction markets (Hanson, 2002). While in traditional prediction markets it is only possible to trade the probabilities for an event $p(A)$ or an event $p(B)$, combinatorial markets also allow to trade the probability of the conditioned events $p(A|B)$ or $p(B|A)$ (and the negated events). The LMSR ensures that, given tree events A, B, and C, an adjustment of the conditioned event $p(A|B)$ does neither change $p(B)$ nor $p(A|C)$ (Powell et al., 2013).

In combinatorial prediction markets each asset is therefore traded by a LMSR and the assets are connected to each other by a Bayesian network (Powell et al., 2013; Sun et al., 2012). Therefore, the law of total probability has to be fulfilled always and an asset has to have the same price if it is traded directly or with its equivalent conditioned probabilities. A trade, however, also leads to the updating of several probabilities and can result in a potentially intractable computational problem if too many relations between questions are defined. Sun et al. (2012) developed and evaluated an algorithm for updating the Bayesian network computational efficient.

DAGGRE and SciCast are two examples of such combinatorial prediction markets. In an experimental setup in the DAGGRE prediction market, it could not be shown that combinatorial prediction markets were clearly superior to flat prediction markets in a "murder mystery" scenario (Powell et al., 2013). It is concluded that the relations between markets have to be very strong to benefit from the conditioned probabilities. Laskey, Hanson, and Twardy (2015) suggest and apply combinatorial prediction markets as a tool to aggregate human (experts) and artificial forecasters. Their prediction platform, named SciCast, is a prediction market to forecast geopolitical events and was operated from 2012 to 2015 by the George Mason University (George Mason University, 2015).

Complexity and Information Overload

Various authors argue that the high complexity of prediction markets and limited understanding of participants are a hurdle to the adaption and use of prediction markets (Teschner and Weinhardt, 2012a). Chen, Li, and Zeng (2015) argue that complexity factors such as dynamic prices, complex pricing mechanisms, and time-dependent and dynamic payoffs pose high cognitive load to its participants that raises entry barriers and lowers intention to participate (especially for non-sophisticated users). According to Chen, Li, and Zeng (2015) this complexity is mainly driven by the selected market mechanism (market micro structure (Weinhardt, Holtmann, and Neumann, 2003)). Therefore, Chen, Li, and Zeng (2015) suggest a fixed odds market design.

Teschner and Weinhardt (2012a) decided not to change the underlying market mechanisms, but the interface. Markets that hide its complexity with easy to use and intuitive interfaces are "Hidden Markets" (Seuken, 2010). Actually all LMSR markets are also hidden markets. In an experimental study, Teschner and Weinhardt (2012a) found surprising evidence that orders submitted with the hidden market interface (trading wizard) were more likely to be profitable, more likely to be market

orders, and smaller in average. It is suggested that complex market interfaces lowers performance by increasing participants' cognitive load.

Yang, Li, and Heck (2015) examined the effect of the information transparency level on different measures in prediction markets. The degree of visibility of orders in the order book also moderates cognitive load. There is potentially a trade-off between the information gain of a transparent order book and information overload. The key findings of Yang, Li, and Heck (2015) state that a *semi-transparent* market, a market where the top three buy and sell orders are displayed, leads to the best accuracy and most interactions between traders. A *full-transparent* market showed no significant further improvement.

Non-observable Events

One disadvantage of most standard prediction market implementations is the modeling of contracts for events with a non-observable outcome (Garcin and Faltings, 2014). According to Sprenger, Bolster, and Venkateswaran (2007) these markets are called conditional prediction markets. In these markets only one of many options will be realized (if at all), such as in the case of prediction markets for idea evaluation (e.g. Soukhoroukova, Spann, and Skiera, 2012), or that the realization may be too far in the future. One approach to solve this problem is suggested by Garcin and Faltings (2014). Contributions to the forecast are rewarded, if later contributions confirm this estimation (simplified: if the price or probability was moved into the same direction). To implement a proper scoring rule in this case, it would be necessary to know the posterior probability distributions of each participant or an additional question that functions as a "Bayesian Truth Serum" (Prelec, 2004; Witkowski and Parkes, 2012). In the case of Garcin and Faltings (2014) latter was replaced by the current poll result. The approach is then called "peer truth serum". Garcin and Faltings (2014) demonstrated comparable performance to prediction markets in the context of Swiss ballots. Therefore, the authors suggest their approach as an alternative to prediction markets where outcomes may not be observable. Further limitations seem to lay in the required existence of poll results (or comparable numbers). It is also not discussed how prone the approach is to manipulation and fraud.

Slamka, Jank, and Skiera (2012) named prediction markets with non-observable events "2nd generation prediction markets" or "G2" markets. These G2 markets just pay off according to the last fixed price as employed by several previous studies (e.g. Dahan et al., 2011; Soukhoroukova and Spann, 2005; LaComb, Barnett, and Pan, 2007). Slamka, Jank, and Skiera (2012) found that this payoff scheme resulted, as expected, in a worse accuracy. However, this was only by 4.4% (topics: sports, politics, and economics). It has to be noted that such markets are, however, more prone to manipulation and fraud.

Manipulation and Fraud

Manipulation and fraud in prediction markets is a very common problem. A detailed investigation is postponed to Section 7.

2.3 Real-Time Delphi

Contents of this section are in part adopted from Kloker et al. (2018c).
See Section A.1 for further details.

2.3.1 A brief History of Real-Time Delphi

The RTD technique is an advanced concept based on the conventional Delphi method, which is already covered by an extensive body of literature. Delphi studies aim to find consensus on questions regarding the future among a panel of experts in a structured and anonymous communication process with feedback. The conventional Delphi method was first introduced by the Air Force-sponsored RAND Corporation in the 1950-1960s (Linstone and Turoff, 2011). The objective was (and still is) to "[...] obtain the most reliable consensus of opinion of a group of experts [...] by a series of intensive questionnaires interspersed with controlled opinion feedback" (Linstone and Turoff, 2002b, p. 10). The first study on the Delphi method was published by Gordon and Helmer (1964) and evaluated the methodology on questions with time-horizons of 10 to 50 years. This publication formed the foundation of a number of Delphi studies in non-defense areas and triggered discussions on the methodology in literature (Linstone and Turoff, 2002b). At the same time the number of conducted Delphi studies also exploded: Hundreds in the 1960s and already thousands in the mid of the 1974s (Linstone and Turoff, 2002b).

In 1975, Linstone and Turoff published a book titled "The Delphi Method" discussing the technique and its applications so far (Linstone and Turoff, 2002b). Today it is considered to be the basic literature for defining the main characteristics and key elements of the conventional Delphi method. According to Rowe, Wright, and Bolger (1991) and Dalkey, Brown, and Cochran (1969) the Delphi method displays four key elements: (1) anonymity of the survey participants, (2) controlled feedback, (3) statistical group response, and (4) iteration.

The basic procedure of a Delphi study is illustrated in Figure 2.4. In a traditional Delphi study, experts receive a questionnaire in paper-and-pencil form from a monitoring team. During a defined time horizon, they give their responses while assessing the specific scenario alternatives by scales of measurement and optionally provide reasons for their estimation. Subsequently, the questionnaires are sent back, and the individual arguments and judgments are summarized by the monitoring team. The group results serve as a basis for a new questionnaire, which is sent back to the respondent group. The experts then have the chance to examine the group feedback and, thereon, reevaluate their original answers. This process can be repeated several times resulting in the occurrence of "structured communication" between the participants, free of "inter-personal" and hierarchical effects, and social pressure (Klein and Garcia, 2015). This makes it especially applicable also for contexts, where different, conflicting interests have to be considered or participants are in an unequal dependency relationship to each other (Linstone and Turoff, 2002b).

Though this basic scheme and the four key principles, the Delphi method was often adapted in order to fit real-world problems and specific contexts (Landeta, 2006). These adaptions also led to the use of the Delphi method in contexts that are not necessarily related to forecasting, such as "budget allocation" or "examining the significance of historical events" (Linstone and Turoff, 2002b). Nevertheless, in forecasting the method is still widely adopted and Landeta (2006) summarizes "[...]

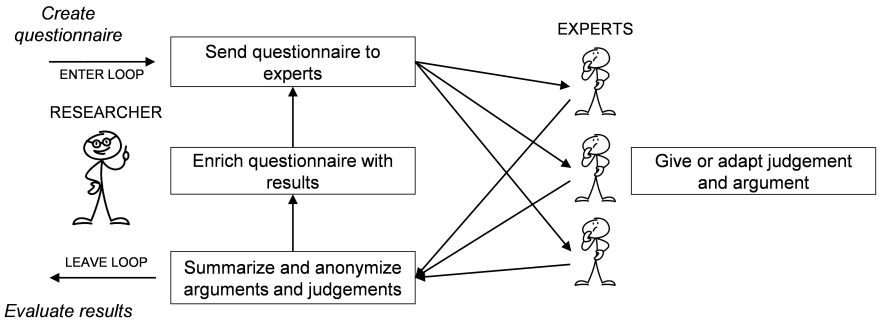


FIGURE 2.4: Process of a traditional Delphi study.

that the Delphi method continues to be used and is a valid instrument for forecasting and supported decision-making" (Landeta, 2006, p. 478).

Rowe, Wright, and Bolger (1991) and Rowe and Wright (1999), however, criticize that a real evaluation of the Delphi method is difficult and leads to a distorted estimation of the quality of the forecasts and conclusions. They advocate that future research needs to elaborate on the mechanics of judgment change within the groups during the rounds and the underlying processes.

Despite the great popularity, the conventional Delphi method also shows some weaknesses, such as the large time frame needed to perform studies, the lack of real-time presentation of results, and the elaborate tasks of the facilitator. The first and second have been addressed by the introduction of the RTD (Gordon and Pease, 2006), empowered by the advanced ICT then available to a broad audience. RTD[16] is an online adaption of the Delphi method that allows participants to access the answers and estimation of other participants not only in the upcoming round, but immediately after submitting (or even before) their own estimation, and, therefore, in "real-time". Access to the study is usually also possible at any time.

Within the last decade, RTD detached itself from the traditional Delphi method and was researched in several studies, though several arguments and definitions still referred to literature of traditional method (Gnatzy et al., 2011; Landeta, 2006). Research on RTDs is based on many studies performed in the last decade. Several publications report the implementation of RTD platforms and present the basic procedures (Abadie, Friedewald, and Weber, 2010; Gary and Gracht, 2015; Schuckmann et al., 2012). In 2011, Gnatzy et al. developed a modified RTD technique based on the idea of Gordon and Pease (2006). The focus of Gnatzy et al. (2011) is laid on the visual statistical group feedback and a higher level of expert guidance through the survey by a one-screen-one-question design.

2.3.2 Real-Time Delphi: Definition and detailed Process

However, current implementations of RTDs differ strongly according to certain aspects regarding the process and the definitions, for which reason they are hard to grasp. This dissertation applies the following working definition:

> *A Real-Time Delphi (RTD) is an online implementation of the Delphi method, where users can interact with the platform online and at any given time.*

[16]The term Real-Time Delphi was already used by Clayton (1997), which meant a Delphi performed during a real meeting. However, this naming was not picked up by other authors.

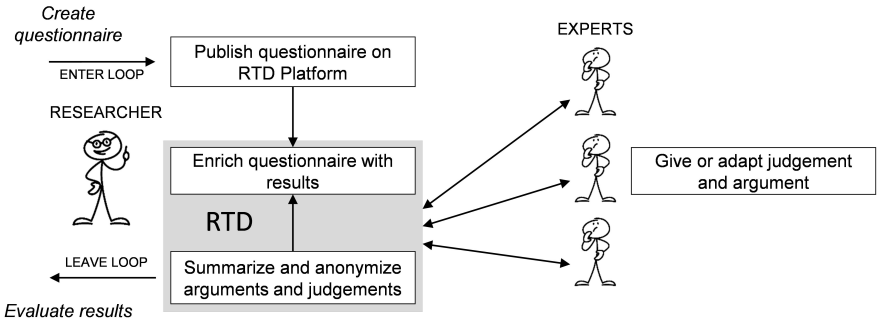


FIGURE 2.5: Process of a Real-Time Delphi study.

A schematic illustration of the process is provided in Figure 2.5. Two central differences distinguishes the process from a traditional Delphi process. First, there are no explicit rounds anymore. Participants are initially invited (and potentially remembered on a regular basis) to access the study which is available online. After their initial estimations, they can decide for themselves, if and when they want access the survey for a second (third, fourth...) time in order to reevaluate new feedback and their own estimations. Second, the feedback is not summarized by the researcher in between the loops, but summarized and displayed to the expert instantly.

However, there is also no clear definition of the process and many variants of RTDs exists (Goodman, 2017).

2.3.3 Ongoing Discussions and Current Developments

Retention

One problem with the traditional Delphi method as well as the RTD is the lack of retention of participants over multiple rounds. According to Mullen (2003) and Walker and Selfe (1996) the response rate in Delphi studies needs to be at least 70%. Reid (1988) notes that the panel size has a strong influence on the drop-out rate. Large panels tend to lead to less retention of participants than small panels with less than 20 members. In RTDs the problem is expected to be much larger, as individual's contribution and involvement becomes smaller. This effect is known as "Social Loafing" and is present in online communities as well (Lampe et al., 2010). Okoli and Pawlowski (2004) argue that the researcher has the possibility to contact the drop-outs and ask them to participate, but this can be, depending on the budget, related to a disproportionate effort (Ishikawa et al., 1993). However, the technological concept of RTD and its asynchronous character would allow distinctly larger panels. To draw upon this potential, it is necessary to bind users strongly to the platform and the survey. This can be accomplished by enabling participants to experience online presence. The experience to feel "present" in the online community is a prerequisite to attribute actions and reactions on the platform to oneself and build up "reputation". Bolger and Wright (2011) found that in "traditional" Delphi studies the promise of gaining social reputation raises motivation to commit to the study and raises retention. To address this issue, the current thesis evaluates a social Real-Time Delphi, described and evaluated in detail in Section 5.

Selection of Experts

The selection of experts for a RTD study is also one key challenge to the method. It often defines, which points are discussed and taken into account and also influences the overall accuracy. Recent approaches suggest to select participants based on their performance in related questions rather than based on reputation (which is common in the traditional Delphi method). Budescu and Chen (2017) build up on the idea that aggregated opinions often outperform individual experts[17] and improve this by selecting those individuals out of the crowd that perform *relatively*[18] better, which means in detail "relative to the crowd's contribution". Therefore, a new measure of contribution is introduced. By this approach of creating a new crowd of *experts* (top positive contributors) Budescu and Chen (2017) shows in two studies the superiority of their proposed model (forecasting current events and economic forecasts of the European Central Bank). This thesis suggests the selection of experts based on their trading behavior in related prediction markets. Details on this approach are outlined in Section 4.

2.4 Overview on selected Research Methods

Within the scope of the presented research several research methodologies were applied, which will be, very briefly, described in this section.

Literature Review A literature review (e.g. Webster and Watson, 2002) intends to analyze relevant research articles in order to gain an extended understanding of the current state and the relationships in a given topic. A guided process helps the researcher to identify all relevant articles given some predefined inclusion and completeness criteria and helps to categorize important common findings and constructs. It is a key step for theory development, as well as to derive research gaps.

CRISP-DM The CRoss Industry Standard Process for Data Mining (CRISP-DM) (Wirth and Hipp, 2000) is a generic process that gives advice on the necessary steps, their order, and implementation of data driven (research) projects. It is regarded as the de-facto standard in data mining projects, though it is relatively high level and therefore allows possibilities for adaptions. The method basically distinguishes between the phases of: Business understanding, data understanding, data preparation, modeling, evaluation, and deployment; which are arranged in an iterative cycle and connected by different uni- and bidirectional sequential paths.

Online Experiments Online experiments (web-based experiments, internet experiments) are experiments not performed in a laboratory facility but dispersed using internet-enabled (private) devices. They are very common in psychology and judgmental forecasting as well (Reips, 2007). Online experiments have the advantage of being more flexible regarding scale and time (and therefore often more cost effective), however provide less control on the subjects and their environments (internal validity). Especially if the subject of matter is an online tool, however, online experiments provide a more natural evaluation environment (external validity).

[17] The wisdom of crowds

[18] In contrast to *absolutely* of traditional weighting models, where the weight is based on absolute past performance (Budescu and Chen, 2017).

Design Science Design Science Research (DSR) (Hevner et al., 2004) is a research method that focuses on the design of innovative artifacts in order to extend human and organizational capabilities, rather than trying to develop and verify theories that explain human or organizational behavior. It bridges the gap from kernel theories from IS, but also other related fields, into the implementation into artifacts of actual use. An iterative process guides the researcher through relevant steps in the research project.

Action Design Research ADR (Sein et al., 2011) is a special form of Design Science and has its focus on projects that are implemented and evaluated in a given organizational context during development and use. Therefore, it meets the requirements of projects that have several organizational restrictions and considers them in the artifact's design and evaluation.

Chapter 3

Delphi-Markets: Integrating Prediction Markets and RTD

> Two are better than one, because they have a good reward for their toil. For if they fall, one will lift up his fellow.
>
> Ecclesiastes 4:9+10, ESV

3.1 Approaches and Potentials

Contents of this section are in part adopted from Kloker et al. (2019) and Kloker, Straub, and Weinhardt (2017a).
See Section A.1 for further details.

3.1.1 Potentials for an Integration

Both judgmental forecasting approaches, Delphi studies[1] and prediction markets, pursue similar goals in many respects. Both methods aggregate many individual opinions (Green, Armstrong, and Graefe, 2007) and can be described as feedback methods (Sprenger, Bolster, and Venkateswaran, 2007). This means that both methods reflect the (aggregated) opinion of the group back to all participants in order to give them the opportunity to correct their own assessments. In addition, both procedures are anonymous (or quasi-anonymous (Kochtanek and Hein, 1999)), making it difficult or impossible to identify other participants. Prediction markets are usually designed as online platforms, a fact that makes the integration with the RTD method particularly interesting. The RTD concept, initially introduced by Gordon and Pease in the year 2006, is an online adaptation of the traditional Delphi method to reduce the duration of surveys. It is characterized in particular by an asynchronous feedback cycle (Kloker et al., 2016). The concept of RTD, which adapts the rigid round-based character of the Delphi method to an asynchronous individualized process, can be easily compared and combined with market participation. A cross comparison of the fields in table 3.1 quickly and clearly shows that the two methods could potentially complement each other in many respects and that some advantages or compensation could be derived from various respective weaknesses.

[1] If Delphi studies are mentioned in the following, it is always referred to an implementation as a RTD. In all other cases the opposite will be explicitly pointed out.

TABLE 3.1: Selected strengths and weaknesses of prediction markets and Real-Time Delphi compared.

	Prediction Markets	**Real-Time Delphi**
Strengths	• Incentives and partially incentive compatible (Jurca and Faltings, 2008) • Gamification (Buckley and Doyle, 2017) • Implicit long-term weighting of participants (Arrow et al., 2008) • Fast reaction to new information (Graefe, Luckner, and Weinhardt, 2010) • Tendency to challenge current forecast (Green, Armstrong, and Graefe, 2007) • Large panel sizes possible (Green, Armstrong, and Graefe, 2007) • Continuous forecast (Graefe, Luckner, and Weinhardt, 2010) • Immediate feedback on estimation (Kranz, Teschner, and Weinhardt, 2014) • Robust against distorted samples (Kranz et al., 2014)	• Collects qualitative and background information • Hypothetical questions and long-time horizons are possible (Linstone and Turoff, 2002b) • More than one question at the same time possible • Questions can be changed after new insights
Weaknesses	• No qualitative information • No background information (and relationships) • Not every question can be transformed in contracts (causalities, hypothetical questions) (Wolfers and Zitzewitz, 2006b) • Observable events are necessary (Slamka, Jank, and Skiera, 2012) • Low visibility for alternative points of view • High complexity (Green, Armstrong, and Graefe, 2007)	• Difficult to achieve incentive compatibility (Green, Armstrong, and Graefe, 2007) • Difficult to motivate participants for long-term participation (Mullen, 2003) • Tendency for conformity (Woudenberg, 1991) • (Only) medium to large sized panels (Vernon, 2009; Green, Armstrong, and Graefe, 2007)

Potentials for Prediction Markets

All information in prediction markets is expressed in price-quantity bundles. For this reason, prediction markets do not permit qualitatively differentiated statements by individual participants, e.g. on the basis of which information and against which backgrounds market participants act and change the market forecast.

By integrating prediction markets and RTDs, participants in a prediction market can be given the opportunity to substantiate their assessment qualitatively or to share information. Basically, there is no incentive to share new information in prediction markets, as these can be used to make a profit. However, the potential gains are usually quickly realized for a single participant, so that participants are willing to exchange it for other information (Kloker, Straub, and Weinhardt, 2017a). In many RTD applications, participants can only see arguments if they participate in the survey themselves. In the sense of reciprocity, but also on the basis of self--expression/self-promotion and prestige (reputation), the exchange of information is stimulated (Kloker et al., 2016). While in classical prediction markets all participants have to acquire information independently of each other, participants in RTD studies learn directly from the other participants. New ideas or scenarios can also be communicated better than via market prices. According to Green, Armstrong, and Graefe (2007), this information exchange leads to a higher efficiency in the information search and also to less information cascades. The latter especially because participants know the background of a price movement and do not have to evaluate and conclude on price movements of which they do not know the important information and the background. A first potential (P1) is thus the improved flow and exchange of qualitative information.

The one-dimensional representation of a question as a stock increases the abstraction load of the participants in the case of complex predictions. Even simple markets are a hurdle for many participants who have little background knowledge about stock trading. Many potential participants lack the necessary knowledge of how expectations are translated into market prices (Green, Armstrong, and Graefe, 2007). This can lead both, to the bounce of participants (with all resulting problems such as distorted samples), as well as to trades that actually does not reflect the participant's expectation properly. The hidden-market approach (Teschner and Weinhardt, 2012a) solves this problem at least partially. In addition, relationships and causalities are difficult to map into contracts, while they are easy to query in RTD studies. A second potential (P2) of an integration with RTD is, therefore, the more flexible design of the prediction object and forecasting question.

Further secondary potentials are, depending on the implementation, a lower number of participants (Abramowicz, 2004), a higher robustness to manipulation (Green, Armstrong, and Graefe, 2007), a better visibility of alternative viewpoints and a higher acceptance of the results by the individual participants (Graefe and Armstrong, 2011). According to Carvalho (2017) the accuracy of a LMSR prediction market may also profit from a round-based participation structure.

Potentials for Real-Time Delphis

In Delphi studies the selection of participants (experts) is the crucial factor for the quality of the results and therefore still a potential weak point (Gordon, 2007; Welty, 1972; Ammon, 2009). For participants without (enough) background knowledge or if all participants draw from the same pool of information, the application of the Delphi method is not advisable or purposeful (Green, Armstrong, and Graefe,

2007; Sniezek, 1990). The selection of experts is also one of the key points that is regularly discussed as a problem to the rigor of the Delphi method (Hasson and Keeney, 2011). In many Delphi studies the participants were selected on the basis of their reputation, which does not necessarily reflect their individual forecasting quality (Hill and Fowles, 1975), and this at often relatively high costs (acquisition, wage/compensation) for participants with a high reputation (Welty, 1972). As a third potential (P3), the prediction markets can be used in integrated approaches for the selection of experts or participants that potentially carry information and or show very high forecasting accuracy. Procedures are described in Kloker, Straub, and Weinhardt (2017a) and Section 3.1.2.

Another problem of Delphi studies is the decreasing motivation of the participants during the process (Cuhls, 2003), since the traditional Delphi process often turns out to be very rigid and lengthy for the participants. Though the RTD approach shortened the process, the problem is still present (as already briefly discussed in Section 2.3.3). The problem results in high drop-out rates, which can be met only conditionally in traditional Delphi studies (Okoli and Pawlowski, 2004; Reid, 1988). Prediction markets offer the possibility to set intrinsic and extrinsic incentives and motivate the participants to a long-term participation, as this results in the largest profits (Green, Armstrong, and Graefe, 2007). The permanent and active participation in the Delphi study can benefit from an integration of these methods (fourth potential (P4)).

Winkler and Moser (2016) deal with cognitive heuristics that lead to systematic errors in predictions in Delphi studies. Persistence of such errors, as, e.g., the anchoring and adjustment heuristic, is demonstrated also by Wilde, Ten Velden, and De Dreu (2018). Winkler and Moser (2016) and Wilde, Ten Velden, and De Dreu (2018) recommend thoughtful creation of proper incentives as a countermeasure. These can be of a financial nature, given in the form of reputation (Winkler and Moser, 2016), or can also be designed as "accountability for the predictions by the participant", so that they have to bear the consequences (Wilde, Ten Velden, and De Dreu, 2018). According to the literal sense of the English proverb "Put your money where your mouth is", participants in prediction markets must also prove the credibility of their arguments with an "investment" or "bet", which can certainly stimulate renewed reflection (Levin, Chapman, and Johnson, 1988). Depending on the underlying market mechanism, a performance-compatible or even an incentive-compatible environment can be created (Chen and Pennock, 2010; Luckner and Weinhardt, 2007). As the provision of arguments may be accompanied by the giving of a trade[2], this may further raise the credibility and confidence of a single argument and provide participants with further information. As a fifth potential (P5), RTD studies benefit from a prediction market through better incentives, reduction of cognitive distortions, and possibly information enrichment of arguments.

A last problematic characteristic of RTD studies, depending on the respective context, is that they are basically designed to find a consensus and thus implicitly suppress disagreement (Green, Armstrong, and Graefe, 2007). As a result, only opinions that lie outside the current consensus are questioned, while a consensus opinion is simply accepted. In prediction markets, profits from a good forecast can be realized only if the own opinion deviates from that of the "crowd", thus the current market price (Green, Armstrong, and Graefe, 2007)[3]. Therefore, every opinion must be questioned and challenged. Nevertheless, in RTD studies the formation of two or

[2] See later the integration approach in Section 3.1.2.

[3] Apart from manipulation and uninformed trading strategies.

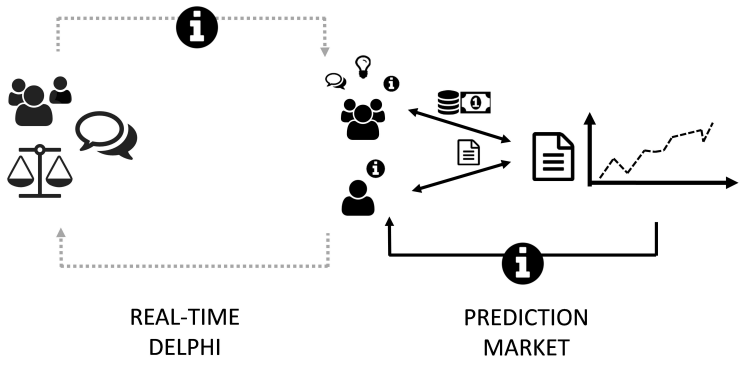


FIGURE 3.1: Integration at user-level. Prediction market and a RTD platform are operated in parallel. User can participate actively on both platforms and possibly transfer information between the platforms.

more opinion clusters is possible (Gnatzy et al., 2011). As a sixth potential (P6), prediction markets offer the advantage that a consensus, the price, is (must be) found even for questions in which different views of the world and values meet and or opinion clusters arise.

Another secondary potential is, depending on implementation and market mechanism, the given incentive compatibility[4] (Chen and Pennock, 2010).

3.1.2 Integration Approaches

However, the term Delphi-Markets, as a term for approaches to integrate RTDs with prediction markets, does not yet make a concrete statement about the versatile possible forms of this integration. Not all potentials mentioned in the Section 3.1.1 can be realized in every integration approach. Three approaches to integration are presented below, to give an impression of the broadness and possible application scenarios. A distinction is made between the integration on a user-, market-, and question-level.

Integration at User-Level

For user-level integration, a prediction market and a RTD platform are operated in parallel. The two platforms are connected via a common user base. This approach is also described in Kloker, Straub, and Weinhardt (2017a) and applied in the FAZ.NET-Orakel. Figure 3.1 illustrates the basic concept. The same participants (or a subset) that trade on the prediction market also participate in the RTD study. Therefore, information of the market flows into the RTD study, while information discussed in the RTD study may become reflected in the prediction market.

The focus of this integration approach is on a panel of prediction market participants. This panel has various features: (1) It includes both informed and uninformed participants (Gruca and Berg, 2007). (2) All participants have a certain interest in the subject (Servan-Schreiber et al., 2004). (3) Participants of the panel possess both public, and private information (Gruca and Berg, 2007). The participants act according

[4] At least for the estimation. Does not apply for the submission of truthful arguments.

to their expectations in the prediction market and are confronted with the expectations of the other participants. For a user-level integration, a RTD survey can now access the participants of this panel. Besides the option to invite all participants from thematically connected markets to the RTD survey, which would only partially solve the problem of selecting experts for the Delphi round mentioned in Section 3.1.1, three interesting alternatives for the selection of participants are conceivable (Kloker, Straub, and Weinhardt, 2017a):

- "The Topscorer": In prediction markets, participants are usually ranked on the basis of their trading success, which according to Hayek (1945) can be achieved only by participants with true (and private) information in the long run. Participants who in the top positions of the rankings are more likely to have important and correct information and thus qualify for the RTD study as experts.
- "The Potential": A problem of the "The Topcscorer" procedure is that a really meaningful ranking for a market can only be calculated after the market has been paid out. For those RTD surveys for which there was no related market recently paid out, experts cannot be selected based on their trading success and their position in the ranking. The selection procedure "The Potential" therefore selects participants based on their trading behavior. It can be shown that certain behavior in prediction markets is correlated with a higher probability of success. Attributes of such behavior can be learned and then used to select experts. Details are outlined in Section 4.
- "The Bohemian": This selection procedure also selects the participants for the RTD survey based on their current trading behavior. The difference to the selection procedure "The Potential" is that not only participants with a high probability of success, but with dissimilar trading behavior are selected. Participants who are more likely to act as buyers may have different information or points of view than participants who are more likely to act as sellers. Against this background it seems sensible to consider both opinions in the Delphi survey and to invite them as participants. In particular, these participants, whose trading behavior appears to represent opinions that differ from those of the mainstream, have potentially new, interesting, or at least previously unnoticed information and points of view that may be of relevance for the Delphi survey. This selection procedure thus leads to a high degree of heterogeneity.

In addition to the selection procedures mentioned above, a random selection would also be possible, since participation in the market already implies a general level of knowledge on the subject. Also conceivable would be a self-selection strategy, according to which the participants themselves decide whether they want to participate in the Delphi survey. According to Green, Armstrong, and Graefe (2007), only these participants participate in the market, who think that their private information has not yet been considered in the previous forecasts.

Actions in the RTD platform therefore have no direct (rule-based) influence on the prediction market or the market forecast. Nevertheless, participants can exchange information between the platforms. In markets, traders usually make profits very quickly from available information, which is why they lose value for the individual trader within a short time (see Fama (1970), Market Efficiency, Section 2.2.1). In order to make further profits from future price developments, traders are constantly dependent on acquiring new information. Among other things, this can be done in exchange with other experts in the RTD survey, which encourages the participants to actively participate in the Delphi survey in the long term.

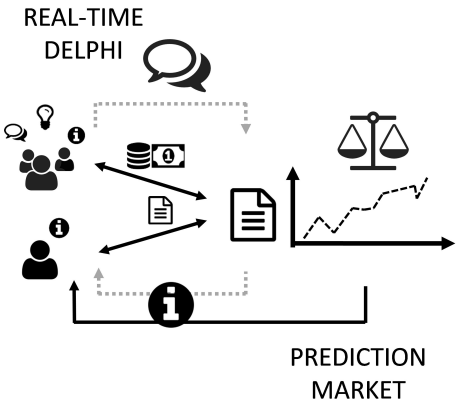


FIGURE 3.2: Integration at market-level. A RTD is performed in the context of a prediction market. Participation in the RTD is not possible without participation in the prediction market.

Another advantage of this integration approach is that the design of the market and RTD can be adapted completely independently of each other to the specific requirements of the respective context. This also applies to the way participants are invited by the market to the RTD platform. Kranz, Teschner, and Weinhardt (2014) recommend here, however, to use an integrated solution, which allows to answer single questions from the Delphi study directly from the prediction markets.

User-level integration thus raises the first and the third potential (P1, P3) and should be used when technical integration is not desired or possible, when conducting the RTD survey is to be decoupled from profit-seeking incentives and trading strategies, or when participants are recruited from multiple pools of people, some of whom have no understanding of markets. In particular, RTD studies benefit from this integration approach especially if the objective selection of participants is of great importance or not possible otherwise (for example due to a lack of historical data on potential participants).

Integration at Market-Level

For market-level integration, a RTD is basically operated within the context of a prediction market. This approach is applied in the same name online platform *delphimarkets.net* and is described in Figure 3.2 conceptually. The figure shows that basically every prediction market is regarded as a RTD study and the submission of an order is also the submission of an answer to the RTD study (and can be accompanied by an argument).

The prediction market with the market question (forecasting target of the market) is at the center of the approach. This is also the first limitation of this approach. Since a market can only answer one question at a time, the RTD survey also consists of only one question. The market acts directly as an aggregation mechanism for the individual opinions of all participants. Prediction market participants have the option to provide the underlying information and background each time they place a trade order. Therefore, each comment is also attributable to a price-increasing or price-reducing effect and thus, in theory, a positive or negative effect on the variable or probability to be predicted. Ensuring the "optionality" of an argument is crucial,

as otherwise many trading strategies aimed at small profits, which are calming the market and providing liquidity, would be prevented. The arguments can then be sorted visually, e.g., as in *delphimarkets.net*, on the sides of the market according to a positive and negative effect on the price. If a participant disagrees with an argument and wishes to write a counterargument, he must necessarily accompany the argument submission of with a trading order. The advantages of this approach are explained in the fifth potential (P5).

This approach has some advantages and disadvantages which have to be weighed up depending on the application. Depending on the market mechanism, integration at market level creates a performance- and or incentive-compatible environment for forecasting the market question, which is why it can be assumed that participants forecast truthfully and with less cognitive heuristics (Wilde, Ten Velden, and De Dreu, 2018; Levin, Chapman, and Johnson, 1988). In addition, this approach allows participants to be intrinsically motivated by the opportunity to share their knowledge (to make a contribution), or extrinsically, by means of a ranking list, by prizes, or even by trading with real money, in the long term. This promotes active and long-term participation in the market and the RTD survey as well as a continuous search for information by the participants (Gangur, 2016). The question about play or real money is thereby subordinated for the forecast accuracy (Servan-Schreiber, 2017)[5]. Another advantage is the user-friendly design of the interaction. Market-level integration, theoretically, allows to display all relevant elements in one view, eliminating the need to switch between different platforms or views. The last advantage of this approach is, as already mentioned in Section 3.1.1, that a market always finds a price, even if different values and opinions are opposed. According to Linstone and Turoff (2002a) the Delphi method is used in particular also for such questions, in which conflicting values and goals have to be considered. A market for finding a consensus price and an argumentation in the style of the RTD for exchanging point of views can complement each other very well here.

A disadvantage of this approach is that although incentive compatibility is achievable for the forecast, this does not necessarily also apply to arguments. In some circumstances, the opposite effect even have to be expected. Participants could use the arguments to lure other participants on the wrong track and profit from the resulting trading orders on the market. Manipulation is a relevant aspect in prediction markets in decisions regarding mechanism and design (Kloker and Kranz, 2017). A further disadvantage results from the fact that prediction markets can actually only be used for questions for which a realization in the foreseeable future is to be registered (Green, Armstrong, and Graefe, 2007). Since in the other case the incentives for providing the truthful estimation are no longer guaranteed, the market price may become object of pure speculative and signaling strategies[6]. Regardless of the time horizon, there can also be other difficulties (complexity or connections between events) that make the formulation of a question into a tradable contract as a market question very difficult or impossible (Green, Armstrong, and Graefe, 2007; Wolfers and Zitzewitz, 2006b). In this context, "Combinatorial prediction markets" (Laskey, Hanson, and Twardy, 2015) offer an interesting option. They allow the modelling of

[5] According to Servan-Schreiber (2017) the choice between play and real money primarily affects the self-selection of participants and thus only indirectly on the forecast(-quality), whereby prediction markets are robust against unbalanced samples in general.

[6] If insiders trade new information, the price moves. This small price movement can be interpreted by other participants as a signal for new information working in a certain direction and, potentially, assessed as credible. This often leads to further price movements in the same direction. This effect can also be used to mislead the market and then make profits from the resulting price movements.

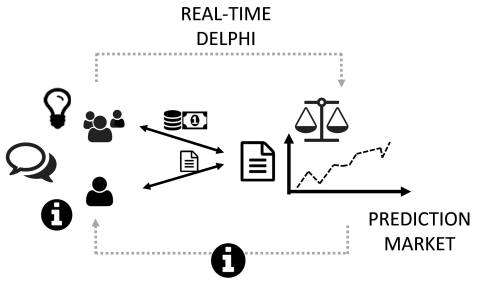


FIGURE 3.3: Integration at RTD-question-level. The aggregation of each Delphi question is carried out by means of a market mechanism.

dependencies between individual markets and would thus address the problem of formulating complex questions as contracts as well as the fact that so far only one market question is possible. However, no implementation or discussion of the integration of a combinatorial market with the Delphi method is known. Finally, the increased complexity of a market is also an obstacle for participants who do not fully understand the market or do not want to deal with it, but would still participate in the RTD study.

Integration at market level therefore raises the first, fourth, fifth and sixth potential (P1, P4-6) and should be used when the issue of forecasting can be represented in one market. The market will benefit from the arguments, while the RTD study will be gamified by the market. However, many Delphi studies will not suit for the break down to one market question, which is why the advantages for the prediction market clearly outweigh within this approach.

At this point, the question arises to what extent this approach differs from a simple (or combinatorial) prediction market with a "forum" or "comment box" for each market question. These are two points in particular: (1) Each comment (or argument) is associated with a price movement, so it is not possible to comment without trading at the same time. In addition, such arguments will increasingly refer to concrete information rather than merely commenting on the question in general or being related to secondary discussions (e.g. fun comments, flaming, etc.). (2) In contrast to simple forums, in RTDs the comments of other participants are usually not visible until an own assessment has been made. In principle, this would also be possible with the trading price (depending on the underlying market mechanism).

Integration at RTD-Question-Level

Basically, for the RTD-question-level integration, a classic RTD is implemented. Only the aggregation of the individual answers to each question takes place on the basis of markets.

The integration is shown schematically in Figure 3.3.

The focus here is on the RTD survey as such and also follows its process. For each individual question, a prediction market is applied in which the participants can trade according their expectations regarding this (forecasting) question. However, there are two problems with this approach: (1) Integration at Delphi question level can become very complex and time-consuming for participants due to the many markets. (2) In addition, the questions in Delphi surveys are usually thematically related and conditional. This also results in markets that condition each other,

which promotes arbitrage and signaling strategies with their potentially negative consequences (truthful response is no longer the dominant strategy).

Hidden markets with scoring rules and combinatorial markets can address these two problems, but only with trade-offs. The use of hidden market can remedy the complexity and time intensity. Depending on the market mechanism used, the Delphi survey with integrated market cannot be distinguished from other Delphi surveys (Teschner and Weinhardt, 2012a; Laskey, Hanson, and Twardy, 2015). If proper scoring rules are used (and this is communicated to the participants), incentive compatibility can also be maintained. Various proper scoring rules are known, the most probably implemented is the Logarithmic Market Scoring Rule (Hanson, 2002) (see also Section 2.2.5). However, the survey partly loses its playful character, which promotes long-term motivation, and also the money metaphor, which could induce deeper considerations (Levin, Chapman, and Johnson, 1988).

However, other advantages remain. This is in particular the long-term implicit weighting of the participants depending on their past forecasting accuracy. Participants with good forecasts increase their portfolio value and thus also their market influence. In contrast, poor long-term forecasters lose their influence on the forecast. Combinatorial markets can prevent arbitrage and signaling strategies, as they can reflect the conditions between the individual markets. However, using combinatorial markets implies several other limits of restrictions on the contract design[7]. Moreover, combinatorial markets are currently only known on the basis of market scoring rules and no other market mechanisms.

An open question that has to be assessed on a case-by-case basis is whether and how the individual markets (for the individual questions) can be fairly combined into an overall ranking without, in turn, promoting arbitrage possibilities and signaling strategies. In addition, a selection of experts by the market, as in P3, is no longer possible.

The integration at the Delphi question level thus raises the first and second potential (P1, P2) from a prediction market's perspective. From a RTD's perspective, more importantly, the approach yields the fourth, fifth, and sixth potential (P4-6). Delphi studies benefit from the long-term motivation of the participants, the additional information supporting the individual arguments, the potential incentive compatibility, the stronger long-term weighting of better participants, and the need for consensus.

This approach is therefore recommended for very complex issues where it can also be assumed that the participants have trading experience (or at least a basic understanding). The variant with hidden markets is recommended if only the former applies, but not the latter, which, however, somewhat weakens the long-term motivation.

Prokesch, Gracht, and Wohlenberg (2015) presented an approach to integrate Delphi studies with prediction markets, which can basically be seen as integration at the RTD-question-level with hidden markets. As a market mechanism, however, a scoring rule was implemented that not resulted in what one would understand under the term "trading". Therefore, the term Delphi-Market is not entirely appropriate and the advantage of classical markets like implicit long-term weighting and the money metaphor are not existent. As in a traditional Delphi study, Prokesch, Gracht, and Wohlenberg (2015) have explicitly selected the participants beforehand. Nevertheless, this approach could beat the relevant benchmarks (Prokesch, Gracht,

[7]Only winner-take-all contracts can be connected within a combinatorial market. Index or spread contracts are not possible, limiting the type of possible questions.

and Wohlenberg, 2015) and shows that Delphi-Markets, applied correctly and in the right place, can compensate the mutual weaknesses of the individual approaches.

3.1.3 Conclusion on Integration Approaches and possible Future Developments

The previous sections have outlined three different approaches to give an impression of possible implementations. The integration on user-level highlighted particularly the potential for expert selection for the RTD study by the market. The integration on market-level shows how a prediction market can benefit from the advantages of Delphi studies (qualitative information). The integration on RTD-question-level took up the potentials for Delphi studies, which result from the market.

Besides these, an integration of the methods may also be possible in other design concepts. However, in most cases, a combination is only effective if an observable event is subject of the prediction, as otherwise the prediction market cannot be paid out and no performance- or incentive-compatible incentives are given. Although there are approaches to the application of prediction markets for non-observable events or events for which no commonly accepted payout value can be determined (Slamka, Jank, and Skiera, 2012), however, these always entail losses in accuracy and raise susceptibility to manipulation (Kloker and Kranz, 2017).

Delphi-Markets are currently still relative rarely implemented and lack methodological research (Prokesch, Gracht, and Wohlenberg, 2015). However, the studies yet known report encouraging results. A significant obstacle and permanent limitations for the use of Delphi-Markets are certainly the increased complexity and the likewise greater implementation effort, which are not justified for every field and case of application. The formalization and validation of the Delphi-Markets as a self-contained research and forecasting approach is still pending and hinders its wide adaption. The variety of possible combinations, however, does not facilitate this task. This dissertation has a focus on the integration on user-level and highlight several points also mentioned in this section.

In general, future research will have to focus in particular on quantifying the theoretically derived potentials, such as the objective selection of experts (P3), or the positive effects of the market structure on the aggregation of expectations for each individual question, as well as the increased (long-term) motivation in RTD studies.

3.2 FAZ.NET-Orakel : Instatiation of a user-level integrated Delphi-Market

In Section 3.1 the integration approach on a user-level was introduced. The FAZ-.NET-Orakel is, as far as this is known, the first tool that combines prediction markets and RTD in a way where participants are shared among both platforms and the prediction market is used to recruit participants (experts) for the RTD study. Many of its features are investigated in detail in the following chapters. This section, however, shall outline the basic project setting and development that put restrictions to the later design and evaluation methodologies. This overview presents all necessary contextual information in order to assess and put later results in their larger context.

The FAZ.NET-Orakel, in general, is a prediction market with all corresponding advantages, but also flaws and problems. Many of the later investigations will reflect this fact. However, during the research project, a RTD platform was fully integrated within the FAZ.NET-Orakel, accompanying several forecasting tournaments,

especially the 2017 German Federal Election. As the RTD enables, unlike as in traditional prediction markets, the "creation of a heterogeneous pool of beliefs", the FAZ-.NET-Orakel can therefore be regarded as a Group Wisdom Support System (GWSS) in the sense of Wagner and Back (2008), who formulated a Design Theory for the class of GWSSs.

3.2.1 Project Setting and Objectives

The FAZ.NET-Orakel is a prediction market running within the online news magazine of the Frankfurter Allgemeine Zeitung, FAZ.NET (http://faz.net). It is a joint project of the Institute of Information Systems and Marketing (IISM) at Karlsruhe Institute of Technology (KIT) (lead), the Frankfurter Allgemeine Zeitung, and the IW Köln. The Frankfurter Allgemeine Zeitung is one of Germany's largest newspapers (0.76 Million Reader, 2017) and the FAZ.NET online magazine likewise one of Germany's largest online magazines (10.29 Million visitors in January 2018) (Wikipedia, 2018). The FAZ.NET-Orakel is the successor of the EIX prediction market that was formerly run in cooperation with the Handelsblatt, another large German online magazine, which was, however, shut down after four years of operation in autumn 2013. While the EIX only featured economic figures and politics, the FAZ.NET-Orakel features markets for politics, economic figures, sports, and other current happenings. The project is online since March 2017. The FAZ.NET-Orakel receives high public visibility, as the FAZ.NET regularly publishes articles on the current forecasting competitions and results. The prediction market is publicly available to all readers of FAZ.NET and beyond, though trading is tied to a FAZ.NET account that can be created free by any holder of an e-mail address. The FAZ.NET is responsible for advertising the platform and is organizing prizes for the tournaments. Until Summer 2018 prizes worth more than €20,000 have been raffled or distributed among the best traders.

The FAZ.NET pursues, besides supporting research, two objectives with the FAZ-.NET-Orakel: (1) Offer their readers an innovative service to engage in order to raise readers engagement and retention and (2) generate own predictions for elections to use and benchmark in news articles. Though the accuracy of the FAZ.NET-Orakel was not evaluated in the course of the research questions underlying this thesis, the FAZ.NET-Orakel regularly predicted events with very high accuracy. Figure 3.4 exemplary illustrates the accuracy of the FAZ.NET-Orakel in comparison with other German forecasting institutions in the context of the 2017 German Federal Election. The FAZ.NET-Orakel was during the complete time between March and October 2017 either the best or at least among the best third of the forecasters. In the "hot period" approximately two weeks before the election, only *Infratest dimap* predicted the result with equal accuracy and some days earlier than the FAZ.NET-Orakel.

Within the university context, the FAZ.NET-Orakel is developed at the IISM at the chair of Prof. Weinhardt in the context of the MInPuD project. The IISM is responsible for the development, operation, and maintenance of the FAZ.NET-Orakel. Objectives of the IISM are to further develop and understand prediction markets and related phenomena. Especially this contains the improvement of prediction markets by integrating RTD studies.

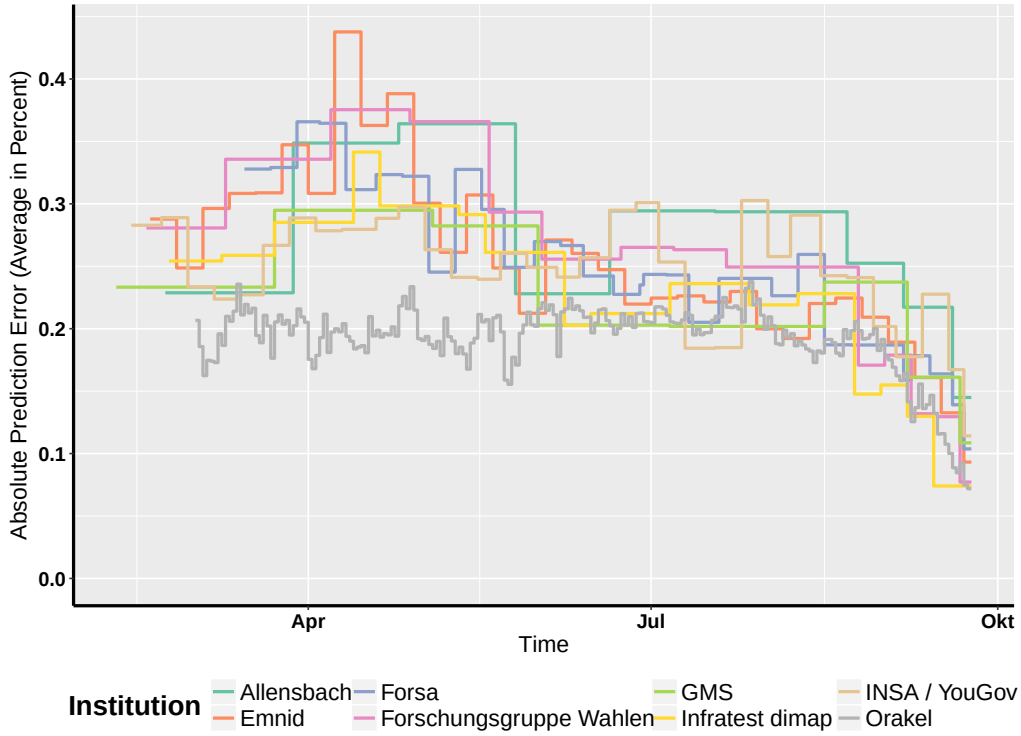


FIGURE 3.4: The FAZ.NET-Orakel in comparison with other forecasting institutions (2017 German Federal Election).

3.2.2 Project Development

The development of the FAZ.NET-Orakel started in the end of 2016 as the contact to the FAZ.NET intensified after some months of informal and non-committal relationship. In December 2016 the basic design of the FAZ.NET-Orakel was adopted to the design of FAZ.NET and approved by the department for corporate design of the FAZ.NET. In January 2017 a one-week test period was conducted for which former participants of the EIX have been invited. Thereafter, the shared login was finalized and at the end of January the necessary contractual agreements were signed. In February and March 2017 the platform was completely debugged and load tests were performed in load generators and with humans in a small lab setting.

The go live for all FAZ.NET readers was in March 2017. The first round of the tournament was running until the end of September 2017, which corresponded to the 2017 German Federal Election. For this period, prizes worth more than €10,000 were raffled and or distributed. As the market itself was operated with play money, no further incentives were given.

Early in March 2017, several fraudulent cases were observed, for which reason a fraud detection algorithm based on Blume, Luckner, and Weinhardt (2010) was implemented. Until August 2017, several other fraud and manipulation preventing elements were implemented. The RTD platform was launched in June 2017 for the Federal Election. After the price round end of September 2017, the end of the second tournament was dated to the end of July 2018, which corresponded to the FIFA Worldcup 2018. In October 2017 the crowd-sourced fraud and manipulation detection tool was implemented and launched. Until February 2018 several small improvements have been added. However, the general interest in the platform and the number of fraudulent attempts dropped distinctly after the Federal Election.

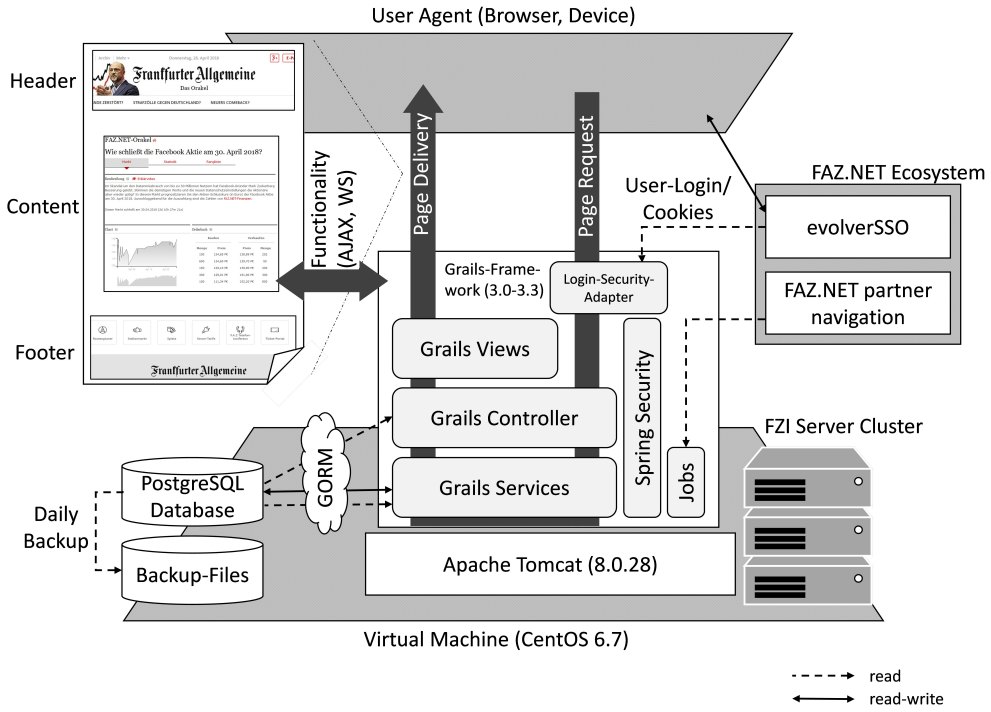


FIGURE 3.5: Software architecture of the FAZ.NET-Orakel.

3.2.3 Technical Perspective

The technical development was completely performed by the author. The FAZ-.NET-Orakel relied on the Java web framework Grails, starting from version 3.0, which was continuously updated during the project to 3.3. Grails is a Groovy[8]-based Model-View-Controller web framework built on top of Spring Boot[9]. Due to the specific requirements of the FAZ.NET, however, several functionalities of the framework had to be enhanced. The architecture is illustrated in a simplified form in Figure 3.5. The back-end was provided in a virtual machine operated by CentOS Linux in the server cluster of the Forschungszentrum Informatik (FZI) in Karlsruhe. A PostgreSQL database was used to store the data. A daily backup ensured data security in case of any downtimes or other potential data inconsistencies. An Apache Tomcat server, installed on the virtual machine, hosted the Grails[10] application to the World Wide Web. The Grails application contained (1) Grails Services, responsible for the data persistence layer and computational tasks, (2) Grails Controller, responsible for the business logic layer and application flow as well as basic data accesses, and (3) Grails Views, responsible for the presentation layer. The sessions were managed by the standard implementation of Spring Security as included in the Grails framework. However, this was extended by a custom Login-Security-Adapter (Filter-Chain-Element) capable to read the session cookie from the FAZ.NET-Ecosystem and manage a shadow-session of the FAZ.NET session in the Spring Security context accordingly.

[8]Groovy is a programming language compatible with and a super set on Java.
[9]Spring Boot is a large Java framework with comparable functionality as Java EE.
[10]https://grails.org

If a user logs in into the FAZ.NET environment, it is assigned a Single Sign On (SSO) cookie from the evolverSSO SSO solution that provides this service in the FAZ.NET ecosystem. At a page request to the FAZ.NET-Orakel, this cookie is compared by the Login-Security-Adapter with session data from the evolverSSO and in case of verification, a shadow session is created in the Spring Security component. This component checks the authorities of the user agent in the component levels Grails Controller and Grails Services. Afterwards, the Grails Controller performs the business logic and decides what will be delivered to the user agent. Computational expensive tasks and write tasks to the database are performed in the Grails Services. This is due to the fact that Grails Services are run as "Singleton Beans" and therefore ensure better data consistency and help to prevent dead-locks on the database. The access to the database is inter-mediated by the GORM data access toolkit that cashes read and write accesses to raise performance. After all data is provided and calculated it is handed to the Grails Views by the Grails Controllers where it is visualized and enriched by the FAZ.NET header and footer. At this point it was also distinguished if the user agent was a mobile device or a tablet/desktop. As the content was fully responsive, only the header and footer of FAZ.NET had to be replaced. Thereafter, the page is delivered. Every further interaction on this page (e.g. order submission, etc.) and all updates on the page (e.g. price chart, etc.) are realized in AJAX (Asynchronous Javascript and XML) and Websockets (WS). Several Grails Jobs are running continuously and call Grails Services to provide important services to the application (match trading orders, fetch the navigation, ...).

Visual features and gadgets, such as the price chart, MicroMarkets (or colloquial: Washer) or the tool for manipulation and fraud detection (see Section 7) were realized in D3.js[11]. MicroMarkets are an additional interface, besides the standard trading form and a trading wizard, to submit one's expectation on a certain event. They accompanied every market and could also be embedded in FAZ.NET articles or other external sites (see Figure 3.6). The expectation has to be submitted by setting two sliders (estimation and confidence). Thereafter, the user has to click "GO" to submit his estimation. A text below the MicroMarket indicates that a trade will be performed based on these settings.

Figure 3.7 illustrates, how stocks, depots, and money is organized in the FAZ-.NET-Orakel. Each forecasting topic consisted of an Accounting Entity (AE), to which an initial amount of money was allocated. To this AE, e.g. the vote-share of the parties in the 2017 Federal Election, several Products were related.

Each product corresponds to one single party and a user can hold stocks of each product. Each product was traded in a CDA with a partly open order book that revealed the best five bid and ask prices (and the respectively available stocks) to the trader. This is close the optimal value of order book transparency determined by Yang, Li, and Heck (2015). Stocks are not transferable between different products and money is not transferable between different accounting entities. This is unlike as in many other prediction markets, e.g. the EIX, where there is "one" money for all markets/products. This design was selected, however, as it ensures that traders who enter the contest at a later point of time, are not disadvantaged in contrast to first-day trader and therefore not discouraged. For each AE an individual Accounting Entity Ranking (AER) in percent was calculated. The best trader received 100%, the worst 0%. All other traders were mapped in between. A total ranking was calculated based on these AERs based on following formula 3.2:

[11] https://d3js.org/

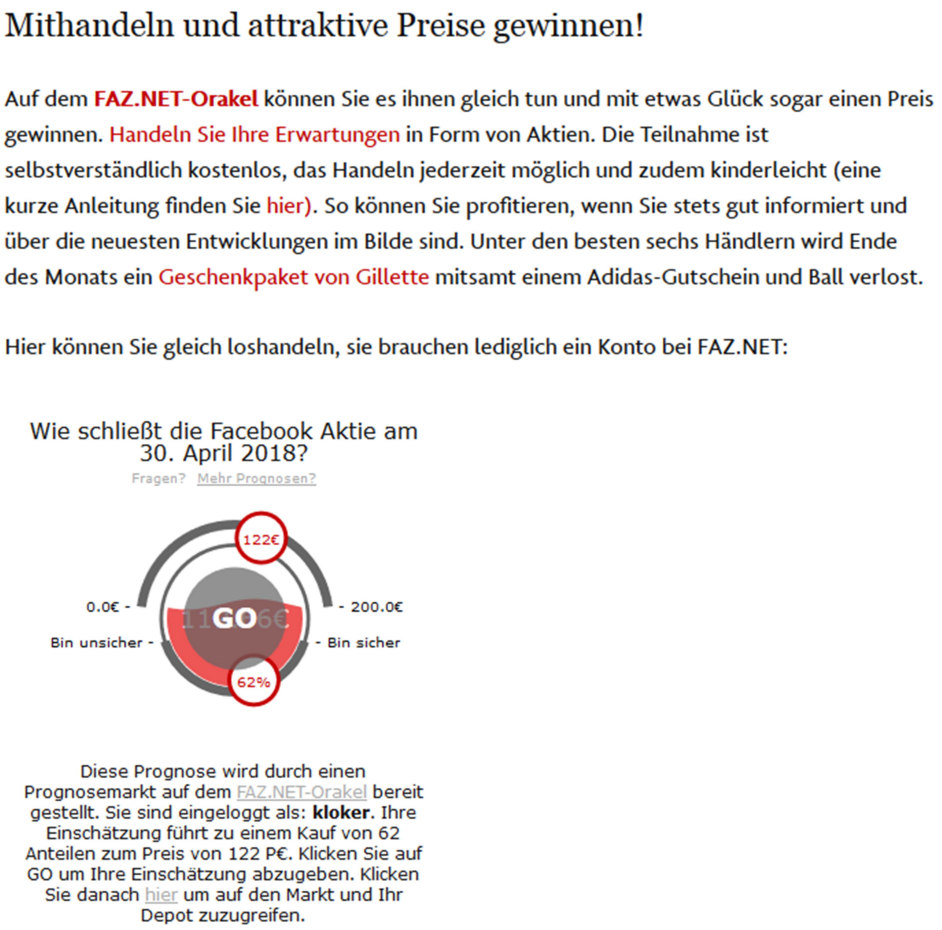


FIGURE 3.6: User interface of a MicroMarket in a FAZ.NET article. As both sliders have been moved, the user is now ready to submit his estimation

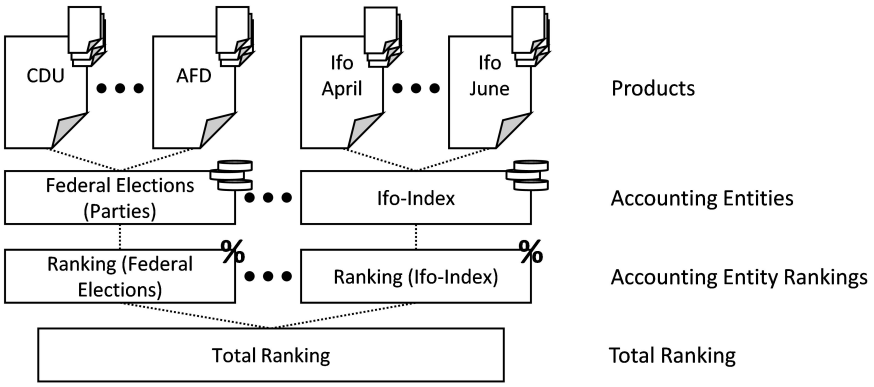


FIGURE 3.7: Concept of AEs and products. AEs bunch together logically related products, so that they can be traded with the same budget (money) and are considered in the same ranking.

$$nom. = count(AERs) + count(arch.\ \ AERs)/2 \tag{3.1}$$

$$score = \sqrt{\frac{avg(AERs) + avg(arch.\ \ AERs)/2}{nom.} * \frac{nom.}{1 + nom.^{0.75}}} * 75 \tag{3.2}$$

The formula ensures that active and long-term participation is encouraged, however, new members have also the possibility to end up in the top positions of the ranking if they perform extraordinary good. During the tournament, old markets were archived and therefore represented in the ranking only with half impact. At the end of the tournament, however, all markets were closed and "archived" which is basically the same as if there had been no differentiation. The last multiplier (75) was dropped in the second round of the tournament.

Chapter 4

Improving Sampling: Select Experts based on Prediction Market Trading Behavior

> You will recognize them by their fruits.
> Are grapes gathered from
> thornbushes, or figs from thistles?
>
> Matthew 7,16; ESV

Contents of this section are in part adopted or taken from Kloker et al. (2018a)
See Section A.1 for further details.

4.1 Problem Formulation

The Delphi method has found vast application for predictions in where expertise from different areas is required, or conflicting goals and values need to be taken into account (Linstone and Turoff, 2002c). However, the Delphi method and its offspring RTD, also face some challenges, such as high drop-out rates or the difficulty of selecting the so-called "experts" (Kloker, Straub, and Weinhardt, 2017a). The aforementioned factors can have a strong influence on the forecast quality and determine which aspects are taken into account and discussed (Welty, 1972; Goodman, 2017). The sampling of the expert panel therefore puts a high risk to the quality of outcome and the rigor of each Delphi study (Hasson and Keeney, 2011). As discussed in Section 3.1, these weaknesses may be addressed by the integration of prediction markets and the Delphi method. It was argued that prediction markets may be used to select the experts (informed traders) to be invited in a RTD survey using an algorithm (of which three were briefly described). The work presented in this chapter is utilizing this idea and aims to define properties of trading behavior in prediction markets that have a predictive power with regard to the trading performance, which is defined as the profit of a trader. According to the Hayek Hypothesis (Hayek, 1945), only traders that are carrying new and valid information will perform

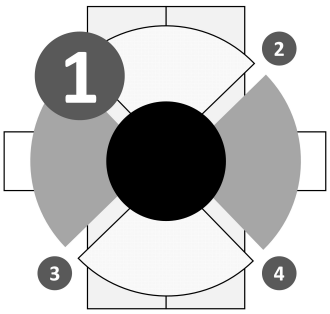


FIGURE 4.1: The presented research project in this section addresses the first source of errors according to the JFIM: Sampling errors.

well in markets in the long term. Comprehensively, these are the experts that should be invited for Delphi studies. However, as it is often not possible to "wait" until the top performers are revealed after the payout of the markets, such a selection has to be performed during the market runtime, when the market result is not settled. Therefore, the approach has to be based solely on trading behavior. Overall, this research addresses following research question:

- Which properties of trading behavior, if any at all, have predictive power on the success of a trader in a prediction market?

The remainder of this chapter is structured as follows: First, the problem of expert selection for RTD studies, as well as the idea of an integration of prediction markets with RTD are briefly picked up in order to understand the problem setting. In Section 4.3, the methodological steps from finding meaningful attributes of trading behavior derived from financial literature to the evaluation of these attributes using regressions and decision trees are introduced. Further, the data used to evaluate these attributes is explained. Sections 4.4 and 4.5 report the implementation and evaluation of the results, respectively. In the final Sections 4.6 and 4.7, the results are discussed and it is concluded that the prediction of informed traders based on the trading behavior is possible to an acceptable extent.

4.2 Related Work

4.2.1 Expert Selection in RTD

Prediction markets and the concept of RTDs have already been introduced in Sections 2.2 and 2.3, for which reason this will be skipped at this point.

Among the challenges that were discussed in Section 3 for the RTD approach, the high drop-out rates and the difficulty to select the experts are among the most noticeable (Teschner, 2012; Welty, 1972; Goodman, 2017; Rowe et al., 2015). The aforementioned factors, especially the expert selection, may have a strong influence on the forecast and may define which points of view are considered (Welty, 1972). Hasson and Keeney (2011) also discusses that the selection of the experts for Delphi studies is in many cases not objective and therefore a real hazard to the rigor of many former and current Delphi studies. Experts are usually selected based on reputation (e.g., based on authorship on academic publications) or based on their profession. However, reputation does not necessarily mean that this person is the most appropriate in regard to the question, as he/she may have a good knowledge about a whole field, but not about the relevant details (Hill and Fowles, 1975). Especially, a selection based on reputation may also cause high survey costs (Welty, 1972). Though, often the selection criteria are cautiously defined and, to a large extent, as objective as possible, many studies still contain the limitation that the expert selection is driven by the experts available and known to the researcher (Rowe et al., 2015). Therefore, it is often not certain if the expert panel's opinion is generalizable to "all" experts. Green, Armstrong, and Graefe (2007) emphasize the problem that the selection of the experts largely influence the topics that are taken into account and discussed in a Delphi study, as well as the overall quality of the results. The method is also not suggested when all experts have access to the same pool of information (Green, Armstrong, and Graefe, 2007; Welty, 1972). Based on these findings, it is to conclude that the expert selection for Delphi studies still lacks a valid and objective methodology.

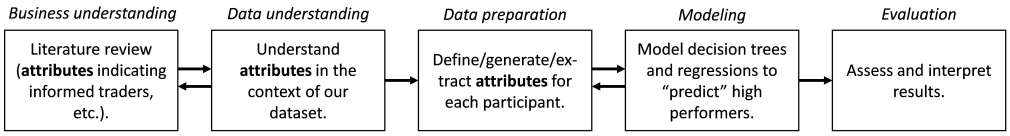


FIGURE 4.2: Research steps to find trading behavior attributes that indicate informed traders. The process is adopted as a single cycle of the CRISP-DM model (Wirth and Hipp, 2000).

4.2.2 Combination of Prediction Markets and RTD to select Experts

Most recently, several publications considered the combination of prediction markets with other forecasting methods. Among others, Atanasov et al. (2016), Tetlock, Mellers, and Scoblic (2017), and Graefe (2015) combined either the forecasts or methods of various forecasting instruments with those of prediction markets or prediction polls, which have led to higher-quality results. Most of them focus on the combination of survey results (opinion polls). Prokesch, Gracht, and Wohlenberg (2015) "combined" prediction markets and Delphi studies, which in this case only meant pre-selecting the traders for a prediction market and using a hidden market interface (therefore, using a market in a Delphi study). One current publication is from Kloker, Straub, and Weinhardt (2017a) (see Section 3.1) and suggests the combination of prediction markets and RTDs on a user-level, so that the market is used for expert selection. Several approaches to this selection are described; the first approach only selects the top performers according to the ranking of the prediction market. A disadvantage is that a valid ranking can only be generated after the market has closed and was paid out. The second approach selects traders with very different trading behavior to select participants with different opinions to enrich the discussion in the Delphi study. A disadvantage is that this does not ensure that the selected traders are real experts. The third approach suggests that an algorithm can be developed that assesses each trader according to his trading behavior for his likelihood to be a high performer, and hence an "informed trader". This approach would mean that those traders are selected as experts that are likely to carry some information about the topic. The current work is suggesting and evaluating the applicability of such an algorithm. An earlier approach to find knowledgeable traders was suggested by Teschner and Weinhardt (2012b) in order to give more impact to their opinion. This approach was based on the price impact of the participants, but could only be calculated afterwards and not during the market run-time.

4.3 Method

To evaluate the trading properties that indicate an informed trader (a high performer), this research follows the process illustrated in Figure 4.2. The process is adapted from the CRISP-DM model[1].

In the first step (Business Understanding), literature from the financial area was consulted. Hypotheses regarding attributes that are, according to this literature, attributed to informed traders, professional traders, institutional traders, or experts in real-world stock exchanges were identified. These attributes are listed in Table 4.1, though there seems to be no consensus on the hypothesis of every attribute. Such

[1] A brief introduction can be found in Section 2.4.

attributes can be of the trader's orders themselves, such as large trading volume (Menkhoff and Schmeling, 2010; Easley and O'hara, 1987), or of the market at the time the trader becomes active, e.g. large spread at the time of order submission (Menkhoff and Schmeling, 2010; Admati and Pfleiderer, 1988). Both properties are, for instance, attributed to a high performer.

Due to the fact that hypotheses for these attributes are from the financial area, it is needed to test these regarding their applicability in prediction markets. Therefore, data of the EIX[2] (Teschner, 2012) was utilized. The EIX is a prediction market that ran for nearly four years, between 2009 and 2013, in four versions in Germany in the context of economic indicators and political elections. In the second step (Data Understanding), previously identified attributes were defined within the context of this data. In the third step (Data Preparation), these attributes were generated and extracted from the EIX data set. The implementation was based on the authors' definitions and are briefly described in Table 4.1.

In the fourth step (Modeling), regressions were performed to check if there are correlations between single attributes and (1) the overall amount of profit and (2) their classification as "high performers". In this context, high performers were defined as the top five percent of each version based on the realized profit. In case (1), a multiple linear Ordinary Least Squares (OLS) regression was applied, showing the impact of each attribute on the amount of profit. As the dependent variable in case (2) is a binary categorical variable, a logistic regression was performed. Both regressions show the impact of the attributes on the dependent variable. In addition, an analysis using Classification and Regression Trees (CART) was conducted. The outcomes are binary trees that can be interpreted as predictive models to classify further data sets. Within a tree, the Gini-index is applied to perform the splits. Using the Gini-index, the algorithm selects attributes for a split based on the homogeneity that may be reached after the split. Therefore, the attributes in the upper part of a tree can be regarded as the attributes that have the best predictive power to subset the participants into homogeneous groups of high performer and other traders.

During the fifth step (Evaluation), it is discussed, if the identified attributes are capable to distinguish between high performers and normal traders. Confusion matrices and two measures, accuracy and precision, were utilized to evaluate the classification trees' performances. Both measures examine the proportion of a tree's correct classifications but in different contexts. While accuracy takes all classifications into account, precision only considers high performers. Comparing the (top) splitting attributes of different trees then led to the attributes that can be applied in an algorithm.

4.4 Implementation

Data of the EIX prediction market was utilized, comprising submitted orders and transactions, aiming to test the predictive capability of the attributes in Table 4.1 and identify the ones to be applied within an algorithm.

The EIX prediction market was run in cooperation with the German online news magazine "Handelsblatt" between 2009 and 2013. Within this period, four versions of the prediction market were running consecutively, which were always changed

[2]Economic Indicator Exchange

[3]As a counterhypothesis of the statement that price-taker are more likely to make errors

[4]An order is considered a round order when mod $10 = 0$ or mod $50 = 0$, in case of an order volume < 50 or ≥ 50, respectively.

TABLE 4.1: Attributes, their hypothetical characteristics according to literature in finance, and implementation.

Attribute	Hypothesis	Implementation	References
market situation (before order submission)			
market activity	informed traders trade in times of high activity	mean orders $+/-$ one hour	Admati and Pfleiderer, 1988
order book size	informed traders trade in times of small order books	median order book size	Menkhoff and Schmeling, 2010; Admati and Pfleiderer, 1988
spread	informed traders trade in times of large spread	mean spread	Menkhoff and Schmeling, 2010; Admati and Pfleiderer, 1988
trader behavior			
order size	institutional traders use mid-sized orders	median order size	Chakravarty, 2001; Barclay and Warner, 1993; Menkhoff and Schmeling, 2010
order size	institutional traders use large orders	median order size	Easley and O'hara, 1987; Menkhoff and Schmeling, 2010
trading volume	informed traders have high trading volume	ordered stocks per day	Menkhoff and Schmeling, 2010; Easley and O'hara, 1987
limit order	informed traders are more likely to place limit orders[3]	prop. of orders not executed immediately	Oliven and Rietz, 2004
trader activity	informed traders participate actively at the market	mean time between two orders	Oliven and Rietz, 2004
round orders	informed traders use round orders	prop. of round orders[4]	Chakravarty, 2001
market maker	experts trade on both sides of the order book	prop. of products with orders offered on both sides	Luckner and Weinhardt, 2008; Harris, 2003
cancel orders	professional trader cancel orders and place them sequentially afterwards	not implemented	Keim and Madhavan, 1995
exper-ience	professional traders learn from previous transactions	not implemented	Luckner and Weinhardt, 2008; Harris, 2003
skill	professional traders have high trading skills	not implemented	Leuthold, Garcia, and Lu, 1994; Aulerich, Irwin, and Garcia, 2013

TABLE 4.2: Different EIX versions and their market activities (only for economic indicators).

Version	Start	End	Contracts	Participants	Transactions
EIX1	10/30/2009	10/31/2010	53	632	22,574
EIX2	10/01/2010	10/31/2011	70	221	11,454
EIX3	11/01/2011	11/15/2012	59	242	11,794
EIX4	11/05/2012	08/14/2013	64	339	16,193

in the fall of each year. For instance, short selling was introduced in 2010, which led to an improvement in prediction accuracy (Teschner, 2012). Table 4.2 shows the four versions of the EIX and some descriptive statistics regarding their market activity. Regarding the forecast accuracy of the EIX prediction market, Teschner, Stathel, and Weinhardt (2011) could show that the EIX forecasts perform competitive to the Bloomberg-survey forecasts, often at an earlier point of time.

Within the EIX data, only the markets for economic indicators were considered, since the field of politics is likely to bring up other traders as high performers than the field of economics. Furthermore, the set of participants was reduced by two criteria: a participant (1) needs to place at least ten orders and (2) has to be active for at least three weeks. This step also leads to the result that the absolute number of high performers in the sample (very unlikely to drop-out in this step) is re-balanced with regard to the absolute number of normal traders. Therefore, this step should also improve the explanatory power of the results. 457 participants remained in the data set and 65 were classified as high performers. Important to realize is that no high performer is lost by this reduction.

To generate the properties for the participants and the market state during each order submit, denoted in Table 4.1, the market was "simulated" by looping and matching the orders in their historical order. However, some attributes are not modeled, as they require to assess the trader with surveys at a later point in time, which was not possible in many cases, and seemed to us as a misappropriate effort for the few cases left.

Based on the traders' realized profits, their classifications, and the generated attributes, regression analysis and methods of CART were applied. In this context, the standard scores of the data set were used to prevent deterioration due to attributes with large variances. Regarding the classification trees, the data set was split into training and test sets by different approaches. Stopping criteria, such as minsplit and complexity parameter, were also manipulated in order to "grow" and compare different trees.

4.5 Evaluation

Table 4.3 reports the results of the regressions. Most notable, one finds that only a few attributes seem to correlate significantly with high profit or the status of a high performer. To be emphasized is the impact of the trading volume, which shows a significant, positive correlation in both: OLS and logistic regression. Moreover, the negative correlation of the order size with the amount of profit is important to realize. The logistic regression further reveals that the attributes spread and market maker seem to be suitable for the distinction among high performers and normal

TABLE 4.3: OLS and logistic regression with attributes, profit, and classification as high performers of EIX versions 1-4.

	Dependent variable:	
	Profit	High performer
	OLS	*logistic*
Market activity	-0.013	-0.228
Order book size	0.001	0.043
Spread	-0.030	0.242*
Order size	-0.126**	-0.141
Trading volume	0.376***	0.569***
Limit order	0.093*	-0.107
Trader activity	0.083	-0.627
Round order	0.054	0.059
Market maker	0.022	0.468**
Observations	457	457
R^2	0.140	
Adjusted R^2	0.122	
Log Likelihood		-145.261
Akaike Inf. Crit.		310.521
Residual Std. Error	0.937 (df = 447)	
F Statistic	8.056*** (df = 9; 447)	
Note:		*p<0.1; **p<0.05; ***p<0.01

traders. All other attributes show no significant impact on the amount of profit or classification as high performer.

In the scope of this analysis, several classification trees were taken into account. Figure 4.3 illustrates one instance of a classification tree that is trained only with the three attributes that showed a significant effect in the logistic regression. The training set consisted of every third trader, starting from the second. Within this tree, the attribute market maker is used as the first splitting criterion. Furthermore, the attributes trading volume and spread are used to classify the market participants. The tree is then used to classify the test set, which in this context comprises all remaining traders. Regarding the performance of this classification tree, Table 4.4 shows the confusion matrix[5] and both measures accuracy and precision. Regarding the classification tree in Figure 4.3, values of 88% and 75% for accuracy and precision, respectively, were achieved. While the accuracy shows that 88% of the trees' classifications are correct, more attention should be given to the precision value. Due to the skewness of the data, there are fewer high performers than other traders. The precision of 75% in this case means that three out of four traders that are selected by the model can be considered as high performers.

By training the classification trees with all attributes and including more as splitting nodes, no better performances is achieved. For instance, in case three additional attributes[6] to those used in the tree in Figure 4.3 were taken into account, a marginally better accuracy of 89% can be achieved, while the precision is lower (69%). Comparing different trees by varying the training and test set, as well as the stopping criteria, draws a consistent picture; the different classification trees mostly make use of the attributes trading volume, market maker, and spread to distinguish

[5] Legend: acc = accuracy, class = classification, cp = complexity parameter, prec = precision

[6] Order book size, market activity, limit order

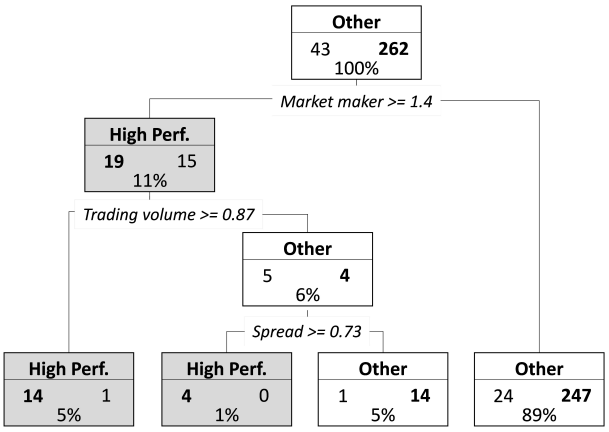


FIGURE 4.3: Reduced decision tree. Trained with the attributes that had been identified in the previous step by logistic regression. Test set is every third participant starting from the second. Training set: The remaining participants (minsplit = 12; complexity parameter = 0.01).

TABLE 4.4: Confusion matrix of the classification tree in Figure 4.3.

		Truth			
		High performer	Other	Σ	minsplit = 12
Class	High performer	6	2	8	cp = 0.01
	Other	16	128	144	acc = 0.88
	Σ	22	130	152	prec = 0.75

between high performers and others. A selection of the best trees is provided in Appendix A.2. This insight is consistent with the results of the logistic regression in Table 4.3.

4.6 Discussion

The results reported in Section 4.5 provide several insights:

Firstly, one can conclude that not all attributes derived from the literature had a significant influence on the traders' results in the ranking. Some, such as the order size, showed a significant effect in the OLS regression, which is, however, reversed to the hypothesis in Table 4.1. This may be due to the fact that the EIX had no transaction costs. Therefore, large orders are not necessary to reduce the proportion of transaction costs and possible revenue. Instead, small orders may allow the participants to better react on fast changing information. Others, such as the trading volume or the market maker attribute, performed as expected. It can be concluded that there may be few differences between trading in financial markets and trading in prediction markets that cannot be further specified on the basis of these results. The lack of a significant effect of some attributes can also be attributed to the fact that the EIX data is much scarcer than financial trading data in stock exchanges, so that some effects might only come to light in cases where more data is available. This can be a limitation at this point, although it has to be discussed whether other and future prediction markets are more "liquid" than the current one.

Secondly, it can be concluded from the logistic regression that there are several attributes with a significant predictive power. These are (large) spread, (high) trading volume, and the market maker attribute. When these attributes are compared with the top split attributes of the classification tree, there is a large degree of agreement. Furthermore, the confusion matrix reported in Table 4.4 shows that a division according to these attributes leads to an acceptable classification. Similar proportions of false and true classifications can be seen in the classification tree in Figure 4.3. Although the classification in both examples suggests that not all high performers were recognized, the number of false positives was very low. This result can be regarded to be a very satisfactory result as it would ensure that the majority of traders invited to the Delphi study are knowledgeable.

Eventually, it is to stress that this result is somewhat surprising, since one have to keep in mind that this classification is only carried out on the basis of behavioral attributes, and not on the basis of information about the traders' knowledge, skills, or experience.

4.7 Conclusion

In view of these results, it can be concluded that a selection algorithm for informed traders or high performers in prediction markets based solely on trading behavior, should be based on these three attributes: (large) spread, (high) trading volume, and the market maker attribute. It was shown that the relatively accurate selection of high performers is possible. Therefore, it is theoretically possible to simultaneously conduct a forecasting market and a RTD survey for the same question/context and to use the prediction market to continuously invite potentially knowledgeable participants as experts to the Delphi study. While the prediction market can then be used as a tool to aggregate information on specific forecasting questions, the RTD survey may be used to collect additional qualitative information and to deal with further secondary questions and or where conflicting values and goals have to be considered (Kloker, Straub, and Weinhardt, 2017a; Linstone and Turoff, 2002c). Since the selection of knowledgeable experts is crucial for the success of Delphi studies (Welty, 1972; Green, Armstrong, and Graefe, 2007), this is an important finding, which opens various possibilities of implementation. The logically subsequent and necessary next step would be to further elaborate on which dimensions expert panels, selected by common methods, differ from panels selected by a prediction market. Furthermore, these findings contribute to the evaluation of the overall integration of prediction markets and Delphi studies by supporting the claim of Kloker, Straub, and Weinhardt (2017a) and Section 3.1.2 that an algorithm for selecting experts on the basis of trading behavior is possible. In addition to this, the presented results may also be helpful for researchers that intend to improve prediction market forecasts by giving experts more weight or try to distinguish informed from uninformed traders to reduce biases, e.g., the favorite-longshot bias (Sobel and Raines, 2003).

When considering these findings, some limitations need to be taken into account. Firstly, it should be noted that the attributes were only tested on one prediction market and context, the EIX and economic indicators, though this still corresponds to more than 60,000 transactions. The second limitation of the current work is that a selection algorithm based on trading behavior is susceptible to misclassification at the start of a market and only becomes more robust during the time of market activity. This corresponds to a cold start problem, therefore, it may be useful to postpone the start of the Delphi study for some time. Next, although supporting evidence was

presented that the selection knowledgeable participants for a Delphi study is possible, it was not shown that this really improves the quality of the result of the Delphi study. This, however, is suggested by theoretical considerations (Prokesch, Gracht, and Wohlenberg, 2015; Welty, 1972). The empirical demonstration of this hypothesis is the next necessary step in future research. Moreover, prediction markets, if they are operated publicly, are subject to a self-selection of participants (Kranz, Teschner, and Weinhardt, 2014). This effect has to be considered in the future research as well.

In summary, it was demonstrated that knowledgeable participants for a Delphi study can be selected by a prediction market. As this was possible with acceptable accuracy during market activity, this work presents a possible solution for a much-discussed problem of the Delphi method (selection of experts (Kloker, Straub, and Weinhardt, 2017a)). In addition, it contributes to the research of prediction markets by comparing the trading behavior of informed traders in prediction markets and financial markets and by demonstrating that there is at least some transferability of hypotheses. Therefore, it is strongly recommended that practitioners implement prediction markets alongside Delphi studies based on current results and other benefits such as long-term user motivation (Kloker, Straub, and Weinhardt, 2017a).

Chapter 5

Improving Response Rates: A social Real-Time Delphi

For, yet a little while, and the coming one will come and will not delay.

Hebrews 10,37; ESV

Contents of this section are in part adopted or taken from Kloker et al. (2016) and Kloker et al. (2018c).
See Section A.1 for further details.

5.1 Problem Formulation

For governments, companies, and organizations, reliable forecasts and assessment of future developments has always been a central success factor (Durand, 2003). These forecasts are generated by statistical models, but also by human judgment. Employees or members within organizations often carry insights and have a gut-feeling about their daily issues that is beyond mere historical and technical data (Styhre, 2002). The RTD methodology is a prominent way to create forecasts based on such "expert panels". However, as many other judgmental forecasting methods, in RTDs forecasts can be distorted by non-response errors (Armstrong, 1985), which are in this particular case a result of drop-outs between Delphi rounds. Such drop-outs (or the lack of retention) are common in Delphi surveys, also because the Delphi rounds often take considerable time (Landeta, 2006). As outlined in Section 2.3.3, retention is a key challenge in RTD and traditional Delphi studies (Mullen, 2003; Walker and Selfe, 1996; Okoli and Pawlowski, 2004; Reid, 1988). Besides for forecasting, RTD is applied in knowledge management, e.g., to develop measurement scales (Boulkedid et al., 2011) or estimate trends and developments (Gnatzy et al., 2011). In all these cases, drop-outs may result in undesired effects on the results, as the panel of participants often defines, which point-of-views and information are discussed and considered in the result or decision (Welty, 1972).

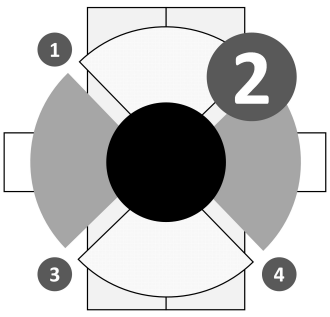


FIGURE 5.1: The presented research project in this section addresses the second source of errors according to the JFIM: Non-response errors.

RTDs are also part of the integrated approach and implemented in the FAZ-.NET-Orakel. In order to design a RTD that solves the design challenge to raise retention, it is first important to understand how a RTD is designed in general. However, a brief look into literature on RTDs shows that different concepts are implemented with often very different designs (e.g., regarding questionnaire layout, process, etc.). Therefore, it is necessary to capture a standard of a RTD (Design Principles (DPs)) that can then, afterwards be extended by a new DP that raises retention. The work demonstrated in the following sections therefore contributes in two ways: First, it provides a summary on all reported implementations of RTDs and derives DPs for a standard implementation of a RTD. Second, it provides the derivation and evaluation of a new DP from literature on retention in online communities. These two steps are achieved in a DSR project.

Overall, this research addresses the following research question:

- How to design social elements in anonymous knowledge sharing and forecasting platforms, in order to stimulate retention in multi-round settings?

The presented results show that "labeling comments" has a significant positive influence on retention. An instantiation of a sRTD artifact is a further contribution of this section's work.

The remainder of this section is structured as follows: Section 5.2 gives an overview of literature on RTD and social interaction on online platforms, to give a foundation to the underlying hypothesis. Section 5.4 introduces the procedure of the DSR project that structured this research. Section 5.5 formulates the DPs. Section 5.6, then, reports the methodology and results of the two-fold evaluation strategy with an online experiment and a field test. A summary of the findings and its limitations, as well as a brief outlook concludes this chapter (Section 5.7).

5.2 Related Work

Real-Time Delphi Research on RTD is based on many studies performed in the last decade. Several publications report the implementation of RTD platforms and present the basic procedures (Abadie, Friedewald, and Weber, 2010; Gary and Gracht, 2015; Schuckmann et al., 2012). However, there is hardly any related work that carries out a systematic literature review to present different approaches of RTDs or discusses similarities and differences between existing implementations. Gordon (2009) refers to other applications, while describing his developed version of RTD. In Gordon, Sharan, and Florescu (2015), this view is enlarged as further techniques are presented. This overview is useful to get an impression of the application areas for RTD, but does not provide any comprehensive comparison of the approaches. In 2011, Gnatzy et al. drew up a modified RTD technique based on the idea of Gordon and Pease (2006). After describing the developed methodology of RTD, Gnatzy et al. (2011) compared their approach to the one of Gordon and Pease (2006). Since Gnatzy et al. (2011) intend to point out detailed improvements of features and process design, the comparison of the implementations is extensive and the adaptions described thoroughly. The focus of Gnatzy et al. (2011) is laid on the visual statistical group feedback and a higher level of expert guidance through the survey by a one-screen-one-question design.

To summarize, there is currently no standard definition of RTD. However, in order to perform a structured literature review and identify DPs (see Section 5.5), there is first the need to define a minimum criteria set to decide whether an application is

a RTD. For this reason the working definition of RTD introduced in Section 2.3.2[1] is applied.

Retention in Delphi and Real-Time Delphi studies One problem with the traditional Delphi method as well as the RTD is the lack of retention of participants over multiple rounds. According to Mullen (2003) and Walker and Selfe (1996) the response rate in Delphi studies needs to be at least 70%. Reid (1988) notes that the panel size has a strong influence on the drop-out rate. Large panels tend to lead to less retention of participants than small panels with less than 20 members. In RTDs the problem is expected to be much larger, as individual's contribution and involvement becomes smaller. This effect is known as "Social Loafing" and is present in online communities as well (Lampe et al., 2010). Okoli and Pawlowski (2004) argue that the researcher has the possibility to contact the drop-outs and ask them to participate, but this can be, depending on the budget, related to a disproportionate effort (Ishikawa et al., 1993). However, the technological concept of RTD and its asynchronous character would allow distinctly larger panels. To draw upon this potential, it is necessary to bind users strongly to the platform and the survey. This can be accomplished by enabling participants to experience online presence. The experience to feel "present" in the online community is a prerequisite to attribute actions and reactions on the platform to oneself and build up reputation. Bolger and Wright (2011) found that in traditional Delphi studies the promise of gaining social reputation raises motivation to commit to the study and raises retention.

5.3 Online Presence to raise Retention in Real-Time Delphi

"Collaboration begins with interaction" (Murphy, 2004, p. 422). The experience of presence in online settings makes geographically separated persons behave as a group, which enriches interaction and the sense of community. "Experience of online presence" is hereby regarded as the degree a participant is feeling personally involved in an online task, which is crucial for forming collaborative communities (Lampe et al., 2010). In a collaborative community, members do not only share perspectives, but are starting to challenge other opinions, reshape their own, and restructure their thinking. This process leads finally to a "shared meaning" – which is also characteristic to the Delphi method. However, online presence in online collaboration has the ability to start additional processes: New perspectives and meanings as well as shared goals can evolve (Roschelle and Teasley, 1995). The second leads to the production of shared artifacts and the intention to "add value" (Kaye, 1992). It is not yet discussed, if these processes lead ultimately to better results in every case, but intuitively one would say that it may improve the result in some dimension. Leveraging this improvement for RTD has not yet happened and Linstone and Turoff (2011, p. 1718) predict that "[...] the future of Delphi will be in collaborative organizational and community planning systems that are continuous, dispersed, and asynchronous."

Online presence and the sense of community allow to build up "social reputation". In order to do this, "labeling" is a widely applied approach in forums. Tagging or labeling (tagging with a fixed set of labels) content in "social question answering" (e.g. Yahoo! Answers or Live Q&A), can open opportunities for richer user interaction (Rodrigues, Milic-Frayling, and Fortuna, 2008). Lampel and Bhalla

[1] A RTD is an online implementation of the Delphi method, where users can interact with the platform online and at any given time

(2007) emphasize the "reputation (status) seeking" behavior of users in online communities. According to Ames and Naaman (2007) there are mainly two reasons to tag social content[2]: i) Providing one's opinion on something (social interaction) and ii) help others/oneself to find something (self-organization). Additionally, Rainie (2007) puts that tagging allows groups to form around points of view and similarities of interest. If persons use the same tags, they may get the impression that they probably share some deep commonalities. Tagging or labeling can therefore contribute to RTDs in multiple ways: First, it enables users to express their opinion about arguments and gain reputation. Second, it enables users to express "common sense". Both leads to a higher experience of online presence and therefore raise commitment to the platform. Third, tagging and labeling are a strong instrument of (self-) organizing content. Especially for larger panels, online discussion can quickly become confusing, if there are no means to structure and distinguish the important from the unimportant or the interesting from the uninteresting. Lots of large online platforms as Twitter, Facebook, Stackoverflow, FlickR, or GitHub use tagging or labeling as a mean to allow the structuring and organizing of content. Turoff et al. (2004) already used labels to organize content in a study which he attested a "Delphi-structure". However, they did not enable the users to label arguments or the inputs of other users, so no social character can be found here. As his panel consisted of students of a lecture and participation was mandatory, also no assumption on retention can be derived here. In addition, his implementation did not fit the anonymity criteria, as names of the authors of arguments were visible. Usually the Delphi method as well as RTD build on absolute anonymity (or quasi-anonymity as argued in Kochtanek and Hein (1999)). Gordon (2009) states the concern about spurious factors, such as (prior) reputation, status, or other social behavior that intrude in face-to-face interactions among experts. These concerns led once to the feature of anonymity in the beginning of the Delphi method. Anonymity is a key feature of the Delphi method and it was adopted in RTD. Therefore, the key challenge of current research is to raise retention by increased online presence and the promise of social reputation with labels and, at the same time, not to allow the tracking of single users long-term and, therefore, harm the anonymity criteria. In the current work and Kloker et al. (2016), generated user names are suggested to support the promise of social reputation and retention. They may induce the feeling of addressability (and therefore the feeling to be present online[3]) and that individual participants may be traced and, therefore, can collect reputation.

5.4 The Design Science Research Project Setting

Aiming at investigating and solving the challenge of low retention by the introduction of social interaction in RTD, the DSR (Hevner et al., 2004) approach as described by Kuechler and Vaishnavi (2008) is applied. Given the current problem and setting, the DSR approach is perceived as promising, because it helps to understand the underlying design and at the same time evaluate an appropriate information system based on the design. As far as this is known, there are no DPs for the class of RTD published yet. The DSR project is conducted in two consecutive design cycles (see Figure 5.2).

[2] In case of Ames and Naaman (2007) photos.

[3] However, this does not mean to perceive the social presence of other, though this was also demonstrated to raise reciprocity in online networks (Teubner et al., 2013).

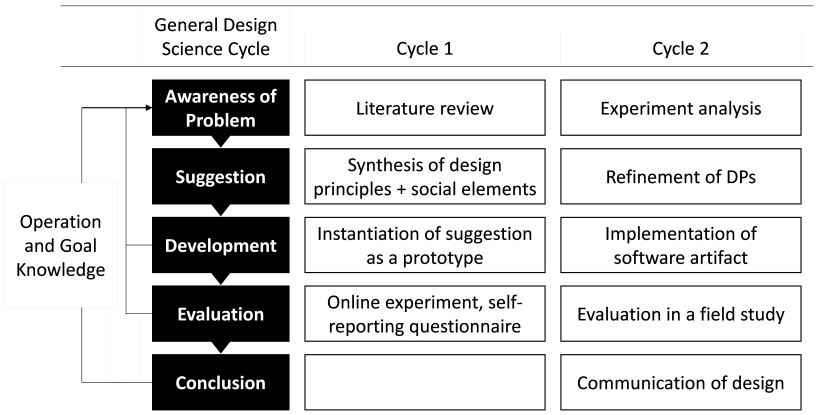


FIGURE 5.2: Two cycle DSR project (adopted from Kuechler and Vaishnavi (2008)).

To create a complete and sound awareness of the problem, the first design cycle was started with an extensive literature review on prior findings in related fields, reported applications of RTDs, and the relevance of the problem with retention in past applications. Based on this review, DPs for the class of RTD platforms were derived. The literature search took all publications into account, in which a RTD approach (or a comparable implementation), based on the working definition of a RTD from Section 2.3.2, was utilized (implemented) and described (or evaluated). Thereupon, a backward and forward search was conducted (Webster and Watson, 2002). Many of the relevant literature was found in the journal "Technological Forecasting & Social Change", in which publishing articles about the Delphi method and RTD had become common. By scanning available descriptions and discussions of different RTD approaches, features were extracted and subsequently organized according to the key aspects of the Delphi method. Thereupon, the approaches were compared according to these features. This allows to differentiate between features that achieved "quasi-standard"-status and features with varying implementations. In the next steps of the first cycle, a preliminary online experiment instantiated and evaluated a prototype of a RTD platform based on these findings in two treatments: (1) A control-group faced a RTD platform implementing the "quasi-standard". (2) A treatment-group faced a sRTD platform implementing the "quasi-standard", plus featuring social elements.

In the second design cycle, the design was reconsidered based on the feedback and results of the first design cycle and the prototype was instantiated into a full IT artifact within and as a component of the FAZ.NET-Orakel. This enabled the evaluation of the artifact under real world conditions. The combination of the literature review, the preliminary online experiment, and the field test together with the industry partner ensures both rigor and relevance of the current research project (Hevner, 2007). Following Venable, Pries-Heje, and Baskerville (2016), a formative evaluation was performed, both ex ante and ex post to determine, how well the artifact achieves its expected environmental utility. The ex ante evaluation can be considered *artificial*, the ex post evaluation as *naturalistic*. In this context the "Human Risk & Effectiveness" DSR evaluation strategy is suitable (Venable, Pries-Heje, and Baskerville, 2016), as the preliminary online experiment is an early formative evaluation of the later naturalistic evaluation in the field test.

TABLE 5.1: Summarized properties of selected RTD studies (Part I).

Publications (→) Properties (↓)	*Studies by Gordon (et al.)* 2006; 2007	2009	2015
Anonymity			
estimations & arguments remain anonymous	✓	✓	✓
Statistical Group Response			
aggregated estimations	✓	✓	✓
average provided	✓	✓	✓
other provided statistics	n/a	median	median, minmax
number of responses	✓	✓	✓
Controlled Feedback			
dissension indicator	flag	flag	n/a
consensus portal	✗	✗	✗
group response is shown ... individual response	before	before	before
others arguments	✓, extra window	✓, extra window	✓
arguments need approval of researcher	✗	✗	✗
visual statistics	✗	✗	✗
Iterative Process			
access anytime	✓	✓	✓
asynchronous process	✓	✓	✓
hot periods	✗	✗	✗
General Properties			
objective of study: find...	consensus	consensus	consensus
initial values in survey	beta test panel	n/a	small pilot group
layout of RTD	2D (matrix)	2D (matrix)	2D (matrix)
depiction of progress	✗	✗	✗
estimation input	4 categories (not linear)	100 and 10 point Likert scale	numeric, multiple choice
argument input	✓, extra window	✓, extra window	✓
internet-based	✓	✓	✓
access to survey by...	user & password	n/a	n/a
integrated tutorial	✗	✗	✗

TABLE 5.2: Summarized properties of selected RTD studies (Part II).

Publications (→)	*Studies by European Business School (EBS)*			
Properties (↓)	Gnatzy et al., 2011	Schuckmann et al., 2012	Darkow and Gracht, 2013	Gary and Gracht, 2015
Anonymity				
estimations & arguments remain anonymous	✓	✓	✓	✓
Statistical Group Response				
aggregated estimations	✓	✓	✓	✓
average provided	✓	✓	n/a	✓
other provided statistics	median, Interquartile Range (IQR)	median, IQR	n/a	IQR, Standard Deviation (SD)
number of responses	✗	✓	n/a	✗
Controlled Feedback				
dissension indicator	color-code	n/a	n/a	color-code
consensus portal	✓	n/a	n/a	✓
group response is shown ... individual response	after	after	after	after
others arguments	✓, extra window	✓	✓	✓
arguments need approval of researcher	✓	n/a	n/a	n/a
visual statistics	✓	✓	n/a	✓
Iterative Process				
access anytime	✓	✓	✓	✓
asynchronous process	✓	✓	✓	✓
hot periods	✗	✗	✗	✗
General Properties				
objective of study: find...	consensus	consensus	consensus	both
initial values in survey	small expert group	small expert group	developed research framework	literature and experts
layout of RTD	1D (1-question--1-screen)	1D (1-question--1-screen)	1D (1-question--1-screen)	1D (1-screen--1-factor)
depiction of progress	in percent	n/a	n/a	n/a
estimation input	metric (0-100%), Likert (5 point)	metric (0-100%), Likert (5 point)	metric (0-100%), Likert (5 point)	metric (0-100%), Likert (5 point)
argument input	✓, directly	✓	✓	✓
internet-based	✓	✓	✓	✓
access to survey by...	hyperlink (e-mail)	hyperlink (e-mail)	n/a	hyperlink (e-mail)
integrated tutorial	✓	n/a	n/a	n/a

TABLE 5.3: Summarized properties of selected RTD studies (Part III).

Publications (→) / Properties (↓)	*Studies by other authors/organizations*		
	Steinert, 2009	eDelfoi	Zipfinger, 2007
Anonymity			
estimations & arguments remain anonymous	✓	✓	✓
Statistical Group Response			
aggregated estimations	✗	n/a	✓
average provided	✗	n/a	n/a
other provided statistics	✗	n/a	n/a
number of responses	✗	n/a	n/a
Controlled Feedback			
dissension indicator	✗	n/a	n/a
consensus portal	dissension portal	n/a	n/a
group response is shown ... individual response	before	n/a	n/a
others arguments	✓	n/a	✓
arguments need approval of researcher	n/a	n/a	n/a
visual statistics	✗	n/a	n/a
Iterative Process			
access anytime	✓	n/a	✓
asynchronous process cmark	✓	✓	
hot periods	✗	✓	n/a
General Properties			
objective of study: find...	dissension	consensus	consensus
initial values in survey	project was prototype	n/a	n/a
layout of RTD	1D (1-screen--1-factor)	n/a	n/a
depiction of progress	✗	n/a	n/a
estimation input	multiple choice	n/a	n/a
argument input	✓	n/a	n/a
internet-based	✓	✓	✓
access to survey by...	hyperlink (e-mail)	n/a	n/a
integrated tutorial	✓	n/a	n/a

5.5 Designing Real-Time Delphi platforms

First, the DPs are derived based on a literature review on RTD platforms. Wagner and Back (2008) defined principles for the class of GWSSs as an extension of Group Decision Support System (GDSS). Following DPs can be understood as a refinement of the principles of Wagner and Back (2008) for the class of RTDs. Literature reports of many applications of RTD studies. However, only few studies reported about the implementation of the method. Probably the two most important "clusters of publications" are those around Gordon and Pease (e.g. Gordon and Pease, 2006) and the EBS, respectively Gnatzy et al. (e.g. Gnatzy et al., 2011). In addition, a cluster of studies are based on the eDelfoi platform (see e.g., Kuusi (1999) or https://metodix.fi). A full comparison of the results is provided in the Tables 5.1 to 5.3 , structured according to the key principles of the Delphi method. Hereinafter, the general findings on an abstracted level in the form of DPs are discussed.

- Design Principle 1: *Ensure anonymity*
  Throughout all publications of implementations of RTD platforms, it is found that the participants remained concealed to the other participants (e.g. Gnatzy et al., 2011).
- Design Principle 2: *Provide meaningful statistical group response allowing self-location*
  Researchers in the field of the Delphi methodology agree that feedback is crucial to achieve results of high quality in Delphi and RTD studies (Rowe and Wright, 1999; Rowe, Wright, and McColl, 2005; Best, 1974). Iterative examination of the group response is essential to find valid consensus on a topic and also to form clusters around alternative positions (Best, 1974). Early RTD studies only displayed basic statistical group response, like the average. Soon many authors started to introduce more meaningful measures as the additional presentation of the median (Gordon, 2009), min-max values (Gordon, Sharan, and Florescu, 2015), the IQR (Gary and Gracht, 2015; Gnatzy et al., 2011; Schuckmann et al., 2012), or the SD (Schuckmann et al., 2012). These measures help the participant to locate his own opinion within the overall range. Less conformity exists regarding the presentation of the number of responses so far. Gnatzy et al. (2011) mention the problem that the feedback pushes participants towards conformity. However, only few authors use RTD without the presentation of the group response, as e.g. Steinert (2009), who explicitly wants to find dissension.
- Design Principle 3: *Use visual feedback to ease the understanding of the statistical group response*
  Closely related to DP 2, diverse authors decided to present the statistical feedback in a visual form (Gary and Gracht, 2015; Gnatzy et al., 2011; Schuckmann et al., 2012).
- Design Principle 4: *Hide feedback before first estimation to avoid anchors*
  This DP is added as it fits more to the traditional Delphi process. Existing implementations show low conformity on this, as only the EBS researchers stick to this principle. However, Gnatzy et al. (2011) states that otherwise the experts become consciously or unconsciously influenced by other participants while forming their own estimation.
- Design Principle 5: *Guide the expert but allow free navigation*
  Literature reports of two different questionnaire structures: 1D and 2D layouts. In the 1D layout (Gnatzy et al., 2011) experts see one question per screen and proceed to the next question by a button. This reduces information overload

and allows to set focus on one question. The order of the questions can be used intentionally, for example to take the expert mentally further in the future by every question. In contrast, in the 2D layout the experts face all questions (in a minimized form) at one glance and can therefore choose the order to answer questions. This is suitable for utility matrixes, decision models, input/output, or, e.g., cross impact (Gordon, 2009). The argumentation must be opened in an external window. Though both implementations exist, the 1D layout is implemented more often. In addition, the 1D instances provide the possibility to navigate to each question directly. So, it is argued that the user should feel guided by the software but can access any question at any time.

- Design Principle 6: *Indicate dissension to highlight where other participants have different opinions*
  Most implementations included an indicator showing the participant that his opinion is out of the group's estimation. This is achieved either by an "indicator flag" (Gordon and Pease, 2006; Gordon, 2007; Gordon, 2009) or by a "color code" (Gary and Gracht, 2015; Gnatzy et al., 2011). Some authors also added a consensus portal, others suggested a dissension portal. Authors that did implement none of the above, usually used a 2D matrix layout that had a comparable function.
- Design Principle 7: *Enable argumentation to allow qualitative discussion*
  Argumentation is a key aspect of the Delphi method. Therefore, this is included in all RTD implementations that are reported in literature. In one case each argument gets reviewed by an administrator before it appears for the other participants. The administrator's task is not to evaluate, but to check the arguments on two criteria: First, eliminate spelling mistakes, and second, delete duplicates to avoid information overload (Linstone and Turoff, 2011).
- Design Principle 8: *Allow access at any time to make the process asynchronous*
  All reported implementations featured an "asynchronous process", so that the access was provided at any time. Gordon and Pease (2006) argue that this comes close to the iterative process in the traditional Delphi method. Not yet "standard" and therefore not a DP, but interesting anyway, is the approach of Kuusi (e.g. Kuusi, 1999) to organize "hot periods", where all participants were additionally invited for specified one or two hour slots to forecast together. This meets the iterative process even better, besides having other advantages (Gordon, 2009): First, this promotes active participation and keeps the discussion ongoing. Second, this ensures that the experts assess the questions simultaneously and also recognize their personal influence shown by immediate reactions of other participants.

With other design decisions that are reported in literature on RTD, either not enough consensus or no necessity to formulate them as DPs was found. These are in particular the realizations of following elements: "deception of progress", "argument input", the "estimation input"; and additional information such as: "confidence", "access to survey", the "end" of the survey (after enough estimation or enough time), and "tutorials and introduction". An important issue, but not a real DP, is that several authors are arguing that the experts should not start with a "null questionnaire" and that initial estimations should be provided by for example a beta test panel etc. (Gordon and Pease, 2006). An alternative is to provide extensive supporting material, definitions, references, or a supportive framework (Gordon and Pease, 2006; Steinert, 2009).

As previously mentioned in Section 5.2, it was assumed that more options for social interaction can raise retention. So a new DP "Enable social interaction to promise the gain of social reputation" is formulated that, however, should be implemented in a way that does not harm anonymity.

- Design Principle 9: *Enable social interaction to promise the gain of social reputation* To raise retention, it is necessary to stimulate experts to a higher commitment to the survey (Linstone and Turoff, 2011). Therefore, it is intended to create stronger bindings to the platform by enabling the experience of online presence. This may be achieved by the increase of addressability or the promise to gain social reputation (Bolger and Wright, 2011). The first one must be handled with caution, as anonymity has still to be ensured. The second may introduce easy to use functionality for feedback. The introduction of, for example, labels that can be added to arguments by all participants, enables each participant to give and receive social reputation (Rainie, 2007).

A RTD, implementing the ninth DP, can be regarded as a sRTD.

5.6 Instantiating the Design

5.6.1 Cycle 1: Prototype Implementation and Evaluation Study

For the first design cycle a prototype was instantiated, implementing design principles 1-5, 7, and 8 that is evaluated in an online experiment. As this was a one shot experiment, DP 6 was skipped, as a dissension indicator first becomes relevant in a second visit.

Method

A two-treatment, between-subject online experiment was conducted, following experimental procedures from experimental economics (Roth, 1986). Therefore, two instances of the platform were set up. From a technical perspective, the online experiment uses a customized web-application following guidelines for online experiments as proposed by Mason and Suri (2012). The first instance implemented a standard RTD setup. The second instance implemented a sRTD and, therefore, also implemented DP 9. It is suggested that DP 9 raise retention in a RTD context.

Stimuli & Experiment Design: Both groups participated in a RTD survey, from which the control group was confronted with the standard version and the treatment group with the social version. The social elements offered together in the sRTD are illustrated in Figure 5.3:

- **Labeling:** The arguments and opinions that have been added to the responses by the participants can be marked with labels such as "good", "bad", or "helpful".
- **User names:** For each question the participants receive a generated user name, which is used to mark their arguments as their own while maintaining their anonymity. These user names can then be used to reference a single person in other arguments or to see which arguments were provided by the same person.

The hypotheses are illustrated in Figure 5.4.

As a proxy to the feature "generated user names", its perception is measured in a *(perceived) Addressability* score, which is defined as the average of 10 items[4][5]. As a proxy to the feature "labeling", its perception is measured in a *(perceived) Promise of Social Reputation* score, which is defined as the average of 9 items[6]. As this was a one shot experiment and there was no additional round, retention could not be measured directly. Therefore, as a proxy to retention, *Commitment* is measured by the average of 3 items[7]. Each item is measured by a five point Likert-Scale. Some items are negated to check for consistent answers. According to the theoretical considerations from Section 5.3 and for the DP 9, it is argued that the *(perceived) Promise of Social Reputation* (introduced by labeling, abbr. "Reputation") has a positive effect of Retention (H_{12}, as shown in Figure 5.4). Based on the considerations in Section 5.3 it is also assumed that the generated user names raise *(perceived) Addressability*, which influences *Commitment* (H_{11}) and *Reputation* (H_{13}) positively.

Participants: The participants were mainly students of Industrial Engineering, Economics, and Business Information Systems from the KIT, Germany. From a contacted 50 participants, 46 answers returned, 22 (female: 9) in the RTD and 24 (female: 10) in the sRTD[8]. For the analysis, only participants that completed both, the survey and the attached questionnaire, were considered. In the RTD and sRTD participants were asked about their opinion on the future of the automobile industry. Therefore, it is expected that the topic does not overly attract the students nor bore them. In addition, due to their technical background, it is to expect that they have at least some expertise.

Procedure: The experiment was conducted in the context of a research seminar. Two students spread the invitation to the RTD survey that was fully functional among their acquaintances via e-mail. The recipients were asked to participate in the RTD/sRTD survey and give their estimations on the future of electronic vehicles. Participants accessed the platform using a personalized link to ensure them ending

[4]Items were taken from Gefen and Straub (2004), Lin (2004), and Rovai (2001). Cronbach's $\alpha = 0.763$ after the deletion of one item.

[5]All items are listed in Table A.2 in the appendix.

[6]Items were taken from Kankanhalli, Tan, and Wei (2005) and Wang and Wang (2010); five self-formulated items were added. Cronbach's $\alpha = 0.786$ after the deletion of two items.

[7]Self-formulated. Cronbach's $\alpha = 0.548$ after the deletion of one item.

[8]A two-sample test for "equality of proportions" (prop.test()) shows no significant inequality (p=.958) regarding the gender.

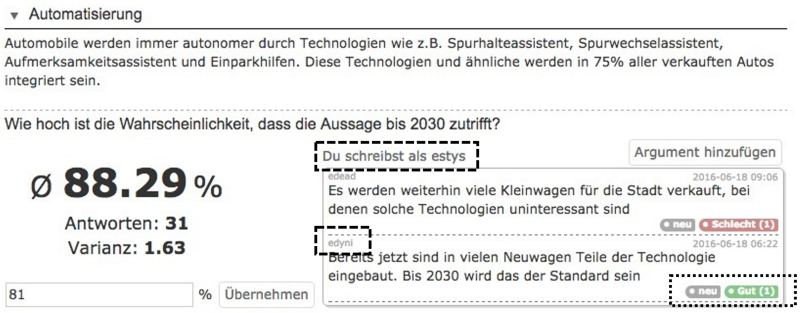


FIGURE 5.3: Social elements as offered to the treatment group (sRTD) in the argument area of a question. Dashed boxes highlight the generated user names in the illustration. The dotted box highlights the labels. Those elements were missing in the control group.

up seeing their own answers when accessing the survey multiple times. Therefore, aborting and re-accessing the survey was possible at any time. The assignment of participants to the treatments was randomized and participants were not aware of treatments. After finishing the survey, the participants received a questionnaire, asking them to self-assess their experience during the survey. The follow-up questionnaire was also conducted with the same platform, however, group feedback and the opportunity to provide arguments were suppressed. The survey was open for one week. After ten days a mail was sent to those participants that participated in the survey and the subsequent questionnaire, saying "thank you" for participation and offered them the opportunity to re-access the platform to see the results, and optionally change their provided estimations. Still, the participants were not aware of the treatments. An event was logged, if they accessed the platform again.

Evaluation of the Prototype

As the gender was equally distributed over control and treatment group, it is not necessary to control for it in the overall effect.

A polyserial correlation between *Commitment* and the treatment ("social" as reference level) showed a significant correlation (r=-.06, p=.039). The data yet reports no significant correlation between the treatment and the revisit after the last e-mail (φ-Coefficient = -.13). The other scores did show a slight, but not significant rise from the RTD to the sRTD treatment. However, a multiple linear regression showed that *Commitment* is, by a considerable portion, explained by *Social Reputation*.

A multiple linear regression was calculated to predict *Commitment* based on *Social Reputation* and *Addressability*. A significant[9] regression equation was found (F(2, 41) = 9.828, p=.000), with a R^2 of .324. Participants' predicted *Commitment* is equal to 0.134 + 0.018(n.s.) *Addressability* + 0.771*** *Social Reputation*, where both independent and dependent constructs were coded or measured on a Likert-Scale from 1 to 5 ("fits not at all" to "fits completely"). A power analysis (Cohen, 1988) with the accepted error levels of $\alpha = \beta = 0.05$ showed that the 46 observations provided enough statistical power (required number of observations > 36). Adding gender as an independent variable results in no significant effect of gender (0.263(n.s.) "female").

[9]Significance codes for all analyses: 0 '***' 0.005 '**' 0.01 '*' 0.05 '.' 0.1 'n.s.' 1

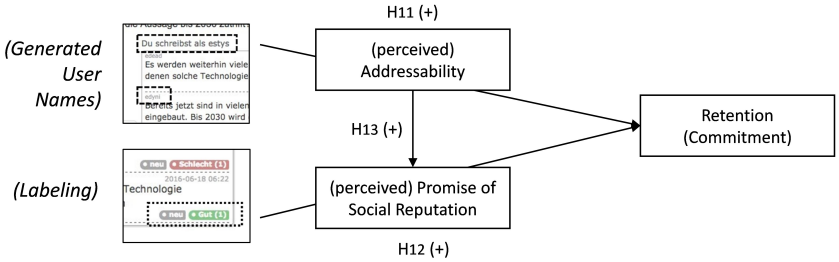


FIGURE 5.4: Research model for the online experiment (prototype evaluation).

A weak positive correlation was, in addition, found between *Addressability* and *Reputation* r(43)=.31* (H_{13})[10].

As an interim conclusion, it can be said that the social treatment led to a significant rise in self-reported Commitment of the participants to the survey and that commitment is by a large portion explained by the self-reported perceived possibility to gain social reputation. A connection between the perceived addressability and commitment, as expected by the design element of the generated user names, could not be shown, though it had a weak correlation with the promise to gain social reputation. Therefore, this design element was dropped in the field test in order to further explore the isolated effect of social reputation on retention. Nevertheless, the feature of generated user names is recommended for future research.

5.6.2 Cycle 2: IT Artifact Implementation and Evaluation Study

For the second design cycle a sRTD artifact was, based on the prototype, instantiated on the FAZ.NET-Orakel and a two round sRTD survey was implemented for the 2017 German Federal Election.

Method

For the evaluation of the Information Technology (IT) artifact, a field study was conducted. The design element of the generated user names was dropped, due to small effects in the prototype evaluation and as it might have confused participants on the FAZ.NET-Orakel, which was not in the sense of the industry partner. Besides, the artifact provided identical functionality than the prototype. The survey was conducted in the context of a prediction market for the vote-share of the parties in the 2017 German Federal Election.

Study Design: The research model for the naturalistic evaluation is illustrated in Figure 5.5. Based on the findings in the prototype evaluation that *(perceived) Promise of Social Reputation* showed a positive effect (see Subsection 5.6.1), the considerations regarding the ninth DP, and Bolger and Wright (2011), it is now hypothesized that *(positive) Social Reputation* increases *Subsequent Platform Engagement* regarding *Retention* (H_1) and *Activity* (H_2). Retention is operationalized as the participation of a user in the second round, if the user participated in the first round. Activity is an additional construct that measures the overall interaction with the sRTD in the second round. It allows a more graduated evaluation of retention. E.g., if all participants from the first round would have participated in the second round, it would still allow to further evaluate the effect of the *(positive) Social Reputation*.

As this was a field test, many other influences may interfere with the dependent construct. Mere increased overall activity could also explain both, (quantitatively) higher *Social Reputation* and *Subsequent Platform Engagement*. Because of restrictions on the platform, no control treatment could be performed and, therefore, it is not possible at this point to finally distinguish between causality and correlation. If a person is very active on the platform in general, this may also explain both: Higher *Subsequent Platform Engagement* and more positive feedback on its arguments. To control for such influences, the overall activity of participants on the platform (in

[10] A test for mediation according to Baron and Kenny (1986) could not confirm mediation, though the results were not finally definite. A singular linear regression from Addressability to Commitment showed no significant effect (H_{11} rejected). However, H_{13} showed a slightly positive effect. For the second reason, a possible mediation cannot be excluded finally.

terms of order submission in the prediction market) was modeled in the proxy construct *Engagement in Prediction Market* as an independent control variable with the hypothesis that the overall increased activity has no effect on the *Subsequent Platform Engagement* (H_3 and H_4) in the RTD survey. No effect of the prediction market as a connected platform on the dependent constructs is assumed, as neither the incentives, nor the objectives were presented as linked between the platforms to the participants.

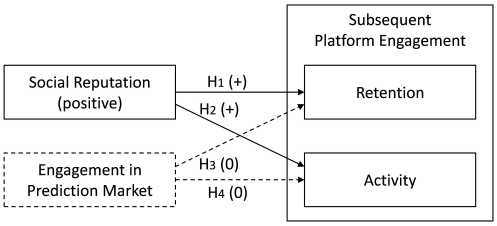


FIGURE 5.5: Research model for field test (artifact evaluation).

From the collected field data, the absolute number of received positive labels before the second round is interpreted as *Social Reputation*, the absolute number of trades in the prediction market for the party results before the second round is interpreted as *Engagement in Prediction Market*, a logical indicator if the participant participated in the second round (either with answers, comments or labels) is interpreted as *Retention*, and the absolute number of answers, comments and labels in the second round is interpreted as *Activity*. Only those participants that participated already in the first round are considered in the evaluation.

Participants: Participants were only recruited among the participants of the prediction market. Here, a certain degree of self-selection towards a knowledgeable group can be assumed, as participation in the prediction market already indicates interest and information on the topic. Participation in the RTD survey was completely non-obligatory and not related to any of the incentives of the prediction market.

Procedure: Technically the software artifact in the field study was identical to the sRTD prototype, besides small design adaptions to the corporate design of the FAZ.NET-Orakel platform. For the field test a survey was implemented asking for the relations between considerations about possible political coalitions and the effect on the expected election outcome for different parties. The survey was always accessible by a link, prominently placed at the top of the corresponding prediction market. Two rounds were performed. Between the two rounds, additional questions (suggestions of the first round by the participants) were added and negatively rated comments were deleted (which is usually the facilitator's task (Gnatzy et al., 2011) and supported by the labels as argued in Subsection 5.3). Round 1 (8^{th} of July to 2^{nd} of August 2017) contained five questions and one question where suggestions could be provided. Round 2 (3^{rd} August to 24^{th} September 2017) added five questions (total: 10 questions). The advertisement for the survey was kept at a minimum: One on-site message at the introduction of the survey and one on-site message at the start of the second round. These measures were taken to ensure that no initial "novelty hype" interfered with the "normal" use of the tool. Some participants have been introduced to the tool itself earlier in a preliminary technical test.

Evaluation of the IT Artifact

Overall, 90 participants participated in the survey by providing answers. Demographic data is not available due to limitations by the industry partner. There is, however, no reason to assume a distinct change of demography between the rounds, for which reason no effect is expected (as also shown in Subsection 5.6.1). 41 participated only in the first round, 7 participants in both rounds, and 42 participants only in the second round (which are, therefore, not considered in the regressions based on first round activity). During the two rounds (first round), 29 (20) comments and 40 (14) labels were provided by the participants. The paths from the independent constructs to *Retention* as illustrated in the research model for the field test (see Figure 5.5) due to its binary coding are tested with probabilistic regressions. The paths to *Activity* due to its continuous coding are tested with linear regressions.

A simple probabilistic regression was calculated to predict *Retention* based on *Social Reputation* during the first round. A weakly significant regression equation was found (p=.093), with (pseudo[11]) R^2 of 0.345. Participants' predicted likelihoods to retain participation is equal to -1.132 + 1.213. *Social Reputation* is measured in the absolute number of positive labels that were received by a participant's comments. A simple linear regression was calculated to predict *Activity* based on *Social Reputation* during the first round. A significant regression equation was found (F(1,26)=5.22, p=.031), with R^2 of 0.167. Participants' predicted likelihoods to participate actively in the second round is equal to 0.723 + 1.058 "content-creating activities".

A simple probabilistic regression was calculated to predict *Retention* based on *Engagement in Prediction Market* during the first round. A not significant regression equation was found (p=.882), with (pseudo) R^2 of 0.001. *Engagement in Prediction Market* is measured in the absolute number of trades in the corresponding markets. A simple linear regression was calculated to predict *Activity* based on *Engagement in Prediction Market* during the first round. A not significant regression equation was found (F(1,26)=5.22, p=.603), with R^2 of 0.011.

To sum up, it can be concluded that participants that received positive feedback in form of labels within the first round had a significantly higher probability to retain in the survey for the second round. It also correlates with a higher activity in the second round. These effects cannot be explained by a general higher activity, as *Engagement in Prediction Market* had no significant correlations with the dependent constructs. Therefore, these results show correlations and suggest causalities in favor of the hypotheses, though the latter cannot be proven finally, due to a non-existent control group.

5.7 Discussion and Conclusion

In the first design cycle, eight DPs for RTD platforms were identified in literature and formulated. The examination of existing RTD platforms showed that mainly two central approaches are established in current research and that they differ especially regarding the survey layout and the moment when the group estimation is presented. In both cases the DPs were formulated in a way that appeared to be closer to the traditional Delphi method.

Subsequently, a ninth DP was added: *Enable social interaction to promise the gain of social reputation*. The need for this ninth DP was derived out of the problem of the lack of retention (Mullen, 2003; Reid, 1988; Okoli and Pawlowski, 2004) during RTD

[11] McFadden (1973)

studies and literature arguing that commitment to online platforms can be raised by social elements (Kloker et al., 2016; Bolger and Wright, 2011; Lampel and Bhalla, 2007).

Thereafter, a prototype of the sRTD was implemented considering all DPs relevant for the preliminary experiment. For the ninth DP the design decisions were based on the suggestions of Kloker et al. (2016) (and Section 5.3), to use individually generated user names per question to maintain anonymity and at the same time introduce addressability. In addition, the possibility to add labels to arguments was provided, so the platform promised users to perceive social reputation when others appreciate their contributions. A two-treatment, between-subject online experiment, and a follow-up questionnaire showed that the assumption that the promise to gain social reputation by social elements raise the self-reported commitment is valid. Though theory suggests, an interaction or positive effect of the generated user names could not be demonstrated. This may have been due to several reasons: First, the generated user names did resemble of vocals and consonants in order to be readable. However, they were not common names and, therefore, may not have induced "social presence" or were recognized as real names for real persons. Second, participants received no explanation, how and for what reasons the user names were generated. Based on the description next to their "own generated name" per question, they may have assumed that each user gets an individual user name per question. However, the experiment design did not check for this certain problem and, therefore, the lack of social presence induced by the user names may be due to a lack of understanding. Nevertheless, as no effect on the dependent variable could be shown, the design element of generated user names was discarded in the field test, though it is suggested for further evaluation in future studies.

The prototype was subsequently instantiated as a full IT artifact and evaluated in a field study. A two round RTD survey was conducted in the context of the 2017 German Federal Election on the FAZ.NET-Orakel. A significant positive effect of the received positive labels in the first round (*Social Reputation*) on *Retention* in the second round and on *Activity* in the second round was found. These effects cannot finally be attributed to causalities, due to the lack of a control treatment in the field test based on restrictions of the industry partner. The research model controlled for the effect that both may be explained by the general activity and engagement of the participants in the overall platform, which could be refused. Therefore, both theory and the results indicate that the introduction of social elements to RTD surveys raise *Subsequent Platform Engagement*.

Limitations of the current work regarding the literature review are especially laid in the bad accessibility of descriptions of many RTD platforms. Hence, only those, which were described in literature, were considered. Limitations of the preliminary experiment are the modest sample size and the questionable reliability of the self-formulated construct *Commitment*, why strong implications should be made cautiously. Future studies may better be based on previously validated constructs, e.g., "IS continuance intention" by Bhattacherje (2001). However, as these items also would have required a reformulation to do justice to the round-base character of the study, it was decided to formulate the items by our-self. Nevertheless, as RTD studies also highlight the asynchronous character, pre-validated measures on general IS continuance and retention independently of rounds may be the better fit for future studies. Limitations of the field test are the lack of a control treatment, which was not possible in the given context on the platform FAZ.NET-Orakel. Therefore, causality and correlation cannot be finally distinguished. It is also to notice that bad comments (labeled by the participants) were deleted between the first and the second rounds by the

moderator to be consistent with traditional Delphi method, where this is usually done by the researchers (Linstone and Turoff, 2011). However, this may also have an effect that cannot be estimated in the current study design. In addition, though a significant effect of positive social reputation on retention was shown, this does not mean that it is just necessary to provide all participants with positive feedback (potentially by the moderator). There is, probably, a certain trade-off between credibility and ludicrousness, which, however, can also not be derived from the current study and should be subject to future research. At last, the field test should be replicated in the future with increased observations and at several points in time to create stronger evidence for the hypothesis and show a stable effect over time.

Nonetheless, the work at hand contributes in many ways to the current research's question, how to encourage participants to take part in knowledge sharing over a long time. It enriches existing literature by formulating the DPs for RTD platforms. In addition, the experiment and field test provided evidence in favor of the expedience of the ninth DP in an experimental setting as well as with real-world data. It was shown that implementing social elements in actual anonymous online settings raised commitment and retention. So, considering the suggested ninth DP helps researchers and practitioners to conduct RTD surveys more successfully, risk less drop-outs, and at the same time raise commitment. This is related with less cost and a higher probability of success. Further research may prove this claim to be true for other knowledge management systems that require subsequent platform engagement.

Chapter 6

Improving Response Quality: Cognitive Factors

> Do not be deceived: God is not mocked, for whatever one sows, that will he also reap.
>
> Galatians 6,7; ESV

Contents of this section are in part adopted or taken from Kloker (2016), Kloker et al. (2017), Kloker, Straub, and Weinhardt (2017b), and Kloker, Straub, and Weinhardt (N.D.)
See Section A.1 for further details.

6.1 Problem Formulation

The long-term success of any company and organization is based on the foresighted and good decisions of its leaders and managers. This involves investment decisions as well as the appraisal of trends and potential challenges. In all of such decision-making and forecasting, however, managers are prone to several cognitive errors (Jones, 2014). Many strategic decisions of managers are based on forecasts on the future development of a certain issue. "Investing" money in a bet on a football team can basically be interpreted as the result of a forecast on the winning team. The same is true for every other investment decision.

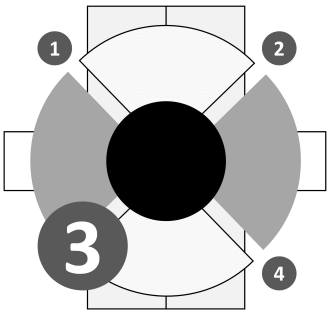


FIGURE 6.1: The presented research project in this section addresses the third source of errors according to the JFIM: Cognitive biases.

In contrast to single expert judgments, group forecasts have the advantage to be less susceptible to certain biases due to the fact that individual differences average out (Winkler and Moser, 2016). Stastny and Lehner (2018) also demonstrated this in a business context, when all participants or experts had access to the same extensive amount of information in form of a report. However, some cognitive biases, such as partition dependence (defined in Section 6.2), still occur systematically in prediction markets (Sonnemann et al., 2011). The occurrences of other (cognitive) biases, such as the home bias, the favorite-longshot bias, or the yogi berra bias are reported by various authors (Luckner and Weinhardt, 2008;

Woodland and Woodland, 2011; Page, 2012). In his popular science book "Thinking fast and slow", Daniel Kahnemann, Nobel memorial price laureate, summarized his lifelong work on the psychology of judgment and decision-making (Kahneman, 2012). In this book, he attributes the occurrence of many biases, such as anchoring and adjustment or the conjunction fallacy, that are often utilized in negotiation strategies and advertisement, to a general inertia of our mind to become active. Kahneman (2012) attributes a certain laziness to the human mind that leads to the fact that often simple "rules-of-thumb" and mental shortcuts, namely heuristics, are applied to many problems and decisions. Therefore, the occurrence of these cognitive biases in human estimations often has nothing to do with an individual's intellectual capacity, but arguably lies with the "slothful nature" of systematic processing (Gómez-Chacón et al., 2014). In this case our mind does not even consider all potentially available and relevant information, but is led by a few heuristic cues that are provided by the context or easy accessible in the memory. In the case of partition dependence, this is the partition of the state space. However, a forecast that is not based on all relevant and available information and is distorted by actually irrelevant information is, in theory, not the best forecast on an event. In the case of partition dependence, it can be demonstrated that the forecast is distorted towards a uniform distribution in a non-rational manner, and therefore misleads forecasters and decision makers. Bardolet, Fox, and Lovallo (2011) showed this effect in real-world investment decisions of managers and Sonnemann et al. (2013) in the case of group-based forecasting.

Dual-Process Theories, which have their roots in psychology and social sciences research, provide a helpful explanation of this phenomenon (Evans, 2012; Watts, 2015). The fundamental idea of the Dual-Process Theories is that there are two different kinds of thinking, of which one is unconscious, fast, automated, and effortless and the other is (at least partially) conscious, slow, and strenuous (induces high cognitive load) (Evans, 2012). Mental shortcuts and cognitive biases are much more likely to happen in heuristic processing. However, several moderators can stimulate or reinforce the use of systematic processing and help to reduce cognitive biases (Watts, 2015). The complexity in which a task is presented is one example of a moderator that can reinforce systematic processing. E.g., in order to create a forecast on the next German Federal Elections, one may just ask people for the vote-share of party XY (expectation polls[1]) or one may invite those people to a prediction market on the vote-share of party XY. In the prediction market the participants first have to translate their expectations into prices and orders and then trade, while the participants in the first group just have to formulate their expectation. Therefore, prediction markets are more complex than prediction polls. However, by this, they can reinforce systematic processing. Zhang et al. (2013) also argues that humans may achieve better results, if they were to consider problems deductively and with more effort. The studies presented in this chapter address the occurrence of the partition dependence bias in prediction markets from a Dual-Process Theories' perspective. Therefore, this chapter puts following research questions:

- Does complexity moderate the partition dependence bias in group-based forecasting tasks?
- Which further factors explain partition dependence in group-based forecasting tasks?

[1] See Section 2.2.5 for an explanation.

To answer these questions, two consecutive online experiments were conducted that both showed significant partition dependence in their estimations. In both experiments, however, the bias did not only occur as a cause of general theory of partition dependence (partition of the state space), but yields strong proneness to other factors. These are especially factors that are reported to be moderators for both, partition dependence (Fox and Clemen, 2005) and heuristic processing according to the Dual-Process Theories (Watts, 2015), e.g. *expertise* and *motivation*.

The remainder of this chapter is structured as follows: Section 6.2 presents selected foundations on partition dependence and Dual-Process Theories. Section 6.3 and 6.4 introduce the methodology, hypotheses, experiment design, and results of the two consecutive online experiments.

Section 6.5 discusses the results and places them within current research, as well as it discusses the limitations. A brief summary of the contributions of this chapter and further research opportunities concludes the section.

6.2 Related Work

6.2.1 Partition Dependence

Partition dependence is "[...] the tendency for the specific partition of the state space to influence judged probabilities" (Sonnemann et al., 2013, p. 11779). This usually leads to a distribution of probability estimations with a tendency towards a uniform distribution (Fox and Clemen, 2005). The roots of research on partition dependence lay in the pruning bias that occurs in "fault trees" (Fischhoff, Slovic, and Lichtenstein, 1978): In the Experiment 5 of Fischhoff, Slovic, and Lichtenstein (1978), it could be demonstrated that persons perceive the likelihood for one branch in a tree to occur lower than the sum of perceived likelihoods for the same branch when split up. To give an example, let us assume two fault trees are designed to find the problem of a car that does not start. A first fault tree may have three branches: Low battery, engine failure, or other causes. A second fault tree may have four branches: Low battery, engine failure, damaged ignition coils, or other causes. The last two branches of the second tree may be fused to the last branch of the first tree. However, experimental evidence shows that if two groups assess the likelihoods for the branches of one tree, in average the sum of the assessed likelihood for the last two branches of the second tree is greater than for the last branch of the first tree. This is also persistent, if descriptions of the events were held constant (if the last branch of the first tree would be called "damaged ignition coils or other causes"). This phenomenon is called the *pruning bias*. Fox and Clemen (2005) generalized this phenomenon to the partition dependence bias. Later, its existence was demonstrated in several other contexts (Bardolet, Fox, and Lovallo, 2011; Tannenbaum et al., 2015; Sonnemann et al., 2013). Bardolet, Fox, and Lovallo (2011) showed that this effect is not only attributable to the fact that people lack of background knowledge in a certain field, but demonstrated partition dependence in investment decisions of managers. With field and experimental evidence and a sample of experienced managers, it was shown that not the characteristics of the companies, but only the number of divisions of the possible options, led to a bias towards the uniform distribution. Fox, Bardolet, and Lieb (2005) suggested that the partition dependence bias can be interpreted as a result of diversification in multiple-item choices and can therefore occur in basically any context and arbitrary tasks. However, Reichelson et al. (2017) argue that this is not a satisfactory explanation of the phenomenon for two reasons: First, the effect could also be demonstrated in single-item choices, where diversification was not possible

(Tannenbaum et al., 2015). Second, Reichelson et al. (2017) report that they were not able to replicate the experiment by Fox, Bardolet, and Lieb (2005) ("candy-bowl task") that was used to support this hypothesis. Therefore, it is assumed that an alternative mechanism besides mere diversification is involved in the candy-bowl task. All in all, Reichelson et al. (2017) argue that there is no support that partition dependence occurs in arbitrary, rather than in conceptually coherent or meaningful, tasks. Although, it is not yet quite clear which mechanisms lead to the partition dependence bias, the bias occurs in a robust manner in judgmental forecasting tasks. In addition, the bias does not "average out" in groups by aggregating many opinions, as many other individual biases do (Winkler and Moser, 2016). Sonnemann et al. (2013) demonstrated in several experiments based on prediction markets the occurrence of partition dependence in forecasting tasks in various contexts. In these experiments, participants were asked to forecast the likelihood that a yet unsettled continuous variable will settle within a certain interval (e.g. the temperature at a certain location and date below or above 0°C). The continuous state space of the possible outcomes was partitioned into three exhaustive and disjunctive intervals. However, two groups of participants made their forecasts on two different sets of intervals. This is schematically illustrated in Figure 6.2, where X is the state space and i are the intervals.

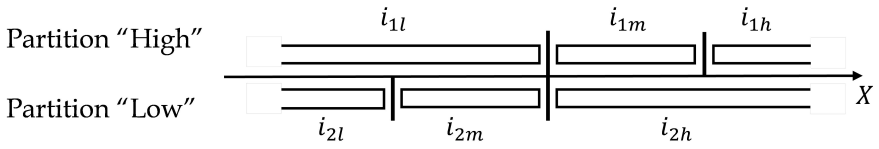


FIGURE 6.2: Schematic partition of the state space and shift between the groups.

The outer intervals (with subscript l or h) are open to the outer sides. In this setting, it is possible to define a measure for partition dependence. The measure of partition dependence pd_x can be calculated by comparing the mean of probabilities assigned to each interval $p(i)$ to the overlapping intervals, e.g. $pd_x = (p(i_{2l}) + p(i_{2m})) - p(i_{1l})$. In the case of no partition dependence it is to assume that the estimated probabilities for i_{1l} should equal the estimated probabilities for i_{2l} and i_{2m}. Partition dependence, however, would suggest a bias towards the uniform distribution, which would lead to the situation that $p(i_{2l}) + p(i_{2m}) > p(i_{1l})$.

Which factors moderate the occurrence of partition dependence (besides the partition of the state space) is only rarely examined. Fox and Clemen (2005) described in their paper that participants with greater substantive expertise show less partition dependence and the effect may sometimes disappear when participants are particularly knowledgeable. However, it is to assume that moderators for other cognitive biases may also apply for partition dependence. Therefore, the next subsection will have a closer look on the concept of the Heuristic Systematic Model, one representative of the Dual-Process Theories, which are usually used to explain cognitive biases.

6.2.2 Heuristic Systematic Model

Previous work on the role of judgment in forecasting, dealing with heuristics and biases, focused on self-assessment (Harvey, 2007). Group decision-making or judgmental forecasting by groups is only rarely discussed. However, the relevance of

heuristics in group-based forecasting is stated by Philip Tetlock, the initiator of the Good Judgment Project. According to his view, "[...] the heuristics- and biases perspective still provides the best first-order approximation of the errors that real-world forecasters make and the most useful guidance on how to help forecasters bring their error rates down" (Tetlock and Gardner, 2015, p. 204). The Good Judgment Project became famous by relatively accurate forecasts on a broad variety of topics utilizing "the wisdom of crowds". In this context also different judgmental forecasting methods were applied and compared, including averaging individual estimations (expectation polls) or prediction markets.

The occurrence of biases is usually explained with the use of heuristics and mental shortcuts during reasoning (Tversky and Kahneman, 1974). When discussing deductive reasoning, the idea of two fundamentally different kinds of thinking in human information processing and reasoning is considered since the 1970 (Evans, 2012) and led to the so called Dual-Process Theories. The basic concept is that there are two different kinds of thinking (Evans, 2012) from which one is unconscious, fast, automated, and easy and the second is (at least partly) conscious, slow, and effortfull (induces high cognitive load). Watts (2015) is summarizing where Dual-Process Theories have been considered in the area of IS. In IS mainly two Dual-Process Theories are considered: The Elaboration Likelihood Model and the Heuristic Systematic Model (Watts, 2015). Both models are variants of the dual-process approach and vary mainly regarding wording (in order to fit certain contexts[2]). The naming of the Elaboration Likelihood Model already suggests its use in the context of information filtering and processing and is often used in Marketing research. The name of the Heuristic Systematic Model is indicating the differentiation between heuristic and systematic processing of information when opinions are built. As forming estimations can be regarded as a subset of the tasks of forming opinions, the Heuristic Systematic Model is a promising approach to explain the occurrence of biases in current context. It also proved applicable in the context of IS (Meservy, Jensen, and Fadel, 2013; Watts, Shankaranarayanan, and Even, 2009). The Heuristic Systematic Model argues that "heuristic cues" trigger heuristic processing. Heuristic cues are design elements or pieces of information that trigger heuristics and discourage people from applying deductive reasoning since it takes more effort. An example of such a heuristic cue is the partition of the state space. For this chapter, the wording of the Heuristic Systematic Model will be applied and it will be differentiated between "heuristic processing" and "systematic processing".

There are several moderators that can initiate or reinforce systematic processing. One such moderator comes in the form of warnings (Winkler and Moser, 2016). Expertise, personality, motivation, or external factors such as incentives, time pressure, or task complexity can be considered as moderators as well (Watts, 2015) . Anything that can influence cognitive capacity or effort can be considered as a possible candidate for the role of a moderator (Watts, 2015) .

The connection between some moderators (e.g., expertise) to the use of systematic processing can be described as "direct". Others, e.g. complexity, operate indirectly by influencing the imposed cognitive load. It follows that it is also necessary to understand the basic concept of the Cognitive Load Theory and how it is related to the Heuristic Systematic Model. First, it is to differentiate between intrinsic cognitive load, extraneous cognitive load, and germane cognitive load (Brünken, Plass, and Leutner, 2003). Extraneous and germane cognitive load can be altered by the

[2]Different Dual-Process Theories usually also use different wording for the systems. Sometimes they also use slightly different understandings of the interactions of the two, or more, types of processing.

design of the instructions, intrinsic cognitive load is a property of the task or material itself (Brünken, Plass, and Leutner, 2003). Put simply, all this cognitive load is distributed on scarce cognitive resources (capacity). When cognitive load matches cognitive capacity we produce better results than if we do not use available cognitive capacity. However, if several cognitive tasks use the same cognitive resources we occur the phenomenon of cognitive depletion (Chen et al., 2017). Basically, this means that our brain is overloaded and we fall back to the use of heuristics and, therefore, are again prone to cognitive biases.

6.2.3 Biases and Complexity in group-based Forecasting

Group-based judgmental forecasting can have various forms (see Section 2.1.2). In this section the focus is on prediction markets. There does exist a great body of work to discuss and explain some of the forecasting errors that appear systematically with the occurrence of (cognitive) biases in prediction markets (Sonnemann et al., 2011; Berg and Rietz, 2018; Cipriano and Gruca, 2014; Cowgill, Wolfers, and Zitzewitz, 2009; Page, 2012; Luckner and Weinhardt, 2008; Woodland and Woodland, 2011). Some examples are the favorite-longshot bias (tendency to overvalue longshots and undervalue favorites) (Snowberg and Wolfers, 2010; Page, 2012), confirmation bias (tendency to ignore conflicting information) (Pouget, Sauvagnat, and Villeneuve, 2017; Cipriano and Gruca, 2014), the overconfidence bias (overestimation of own skills or results) (Berg and Rietz, 2018), or – subject of this research – the partition dependence bias (Sonnemann et al., 2013). There are different explanations why each bias occurs based on market mechanism, market liquidity, information spread, or motivation of participants[3]. However, yet there is not a clear explanation for all biases in prediction markets, but Winkler and Moser (2016) showed that warning messages reduce some biases significantly. This already suggests that they happen during the information processing and estimation formation of each individual (Winkler and Moser, 2016). In prediction markets, events with multiple outcomes have to be presented as a set of different stocks. These different stocks are basically a division of the state space, for which reason prediction markets are very prone to this bias.

Besides the complexity of the topic (intrinsic cognitive load), the overall imposed cognitive load of the forecasting task is moderated by the manner and mechanism in that the estimations needs to be processed to be expressed (germane cognitive load) (Chen, Li, and Zeng, 2015). Chen, Li, and Zeng (2015) argue in that the underlying (market) mechanism has a huge influence on the cognitive load imposed on the participant. Chen, Li, and Zeng (2015) classify different market mechanisms for prediction markets according to their imposed cognitive load/complexity. They used following dimensions for their classification: Pricing (auction vs. posted price), timing (dynamic price), revisiting (dynamic price), and benefit (dynamic payoff). According to Vakkari (1999), a task the complexity of a task rises, the more and longer alternative paths to the solution exist and the less the solution is determinable. In the case of prediction markets, a forecast cannot only be provided as the expectation itself, but has first to be translated into prices. Then a strategy has to be found that maximizes the expected payoff of trades given one's expectation (for the moment

[3]For example, the favourite-longshot bias is attributed too low liquidity that leads to the situation that in the border areas orders do not get matched, which can again be explained using the prospect theory (Snowberg and Wolfers, 2010). Other explanations argue with the composition of the field of participants (Feess, Müller, and Schumacher, 2014; Restocchi et al., 2018), the motivation of the participants, and subsequently incentives (Servan-Schreiber et al., 2004).

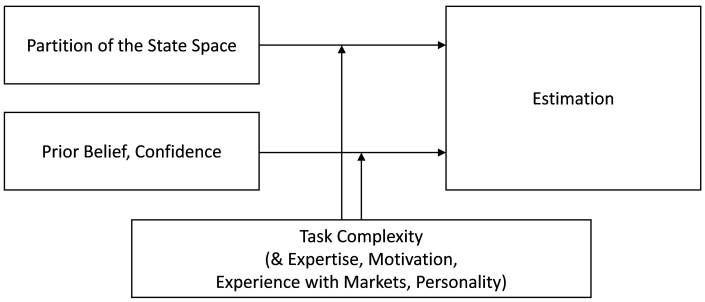


FIGURE 6.3: Basic research model for experiment 1 and 2 (based on Watts (2015)). Complexity and other moderators' influence to which extent the heuristic cue (partition of the state space) affect the final estimation.

and possibly for the future). Participating in a prediction market based on a CDA is therefore "very" complex, while placing a bet on a betting page with predetermined odds is less complex (Chen, Li, and Zeng, 2015). Markets featuring a market scoring rule (e.g. the LMSR, see Section 2.2.5) are less complex than markets featuring a CDA, but still more complex than posted-price markets (Chen, Li, and Zeng, 2015). In markets with scoring rules, participants do not trade against each other directly, but against an algorithm that adapts its offered price according to the bets of earlier participants, minimizing his own risk. If considerations regarding manipulation, signaling, or other strategies to take influence on other participants were left aside, a dominant strategy is to buy or sell until either a participant has no more means, or the offered price matches a participant's expectation. Therefore, the translation of an expectation into a price is much easier in markets with a market scoring rule than in markets featuring a CDA.

6.2.4 Fusion of the Related Work into a Basic Research Model

Summarizing related work, the following research model can be formulated (adapted from the standard Dual-Process Theories research models in IS from Watts (2015)). The research model is illustrated in Figure 6.3.

It can be assumed that an estimation on a future event is optimal, if it represents all the information that is available to the participants in an unbiased form. However, heuristic cues, such as the partition of the state space, distort this information. In case of partition dependence, the estimations are biased towards a uniform distribution. The more a person is using his systematic processing, the effect of the heuristic cue to distort the estimation is assumed to be decreased. The degree of systematic processing is moderated by the imposed cognitive load. From this it is derived that a more complex setting should stimulate systematic processing and reduce heuristic processing. Further moderators that may influence the occurrence of the partition dependence bias are *expertise* as suggested by Fox and Clemen (2005), *motivation* and *personality traits* as mentioned by Watts (2015), and *experience with markets*.

While expertise, motivation, or experience with markets are properties of the individual, complexity can be manipulated by the market mechanism. Based on the previous section, it can be argued that a market featuring a LMSR is more complex with regard to the definition of the sections above and puts more cognitive load on a

participant than just providing one's expectation in an expectation poll. According to the Heuristic Systematic Model, it can be assumed that the occurrence of biases should be less in the market with a LMSR (within certain boundaries). It is also assumed that this is true for the partition dependence bias. In the classification of Chen, Li, and Zeng (2015) this would mean to alter the pricing from "fixed" or "no pricing" to "dynamic pricing", and, therefore, the LMSR market comes along with higher cognitive load. This two *treatments* have the advantage that the interfaces can be designed very alike.

However, especially the effect of complexity is constrained by cognitive depletion. Based on the Cognitive Load Theory and the phenomenon of depletion, theory suggests that participants that are not used to markets may be over-strained and fall back to heuristics. Meub and Proeger (2016) showed that complex forecasting tasks (which induced high cognitive load) raised the susceptibility to cognitive biases (anchoring) even in presence of monetary incentives. In addition, Sonnemann et al. (2013) demonstrated the bias in prediction markets with continuous double auctions, which are quite complex. The current work suggests to compare prediction markets with a LMSR and simple expectation polls (Rothschild and Wolfers, 2013) in an experiment, assuming that LMSR markets do not result in cognitive depletion.

6.3 Experiment 1

This preliminary experiment was designed and used to understand the relationships between complexity and partition dependence, the applicability of the Heuristic Systematic Model, and other influencing drivers better.

6.3.1 Method

Stimuli & Design

The experiment consisted of three forecasting tasks on economic figures (DAX-30, Diesel price, Deutsche Bank stock). Each economic figure can settle on a continuous state space. This state space is divided two times into three disjunctive and together exhaustive intervals, in such a way that the outer intervals are open to the outer side and that they are shifted by the size of the inner interval (see Figure 6.2). Doing so, one receives for every economic figure in the experiment a "high" and a "low" partition (see Figure 6.2 and Table 6.1). Participants have to assess the probability for each interval that the realization of the economic figure will fall into it. For the intervals belonging to the same partition of the state space this probability has to sum up to 100%. For each task, every participant faces either the low or the high partition, never both. Therefore, partition dependence can be measured for each forecasting task, as explained in Section 6.2. This experiment design is adapted from Sonnemann et al. (2013), though they used markets featuring a CDA.

In experiment 1 three treatments are defined:

- In the first treatment the participants assessed the probabilities[4] for each task and interval in an expectation poll. This treatment is referred to as *poll*.
- In the second treatment the participants assessed the probabilities for each task and interval in a market featuring a LMSR. Each participant was alone on this

[4]The probability that the realization of the event will fall into this interval.

market. Therefore, the "initial" prices shown to the participant equaled the uniform distribution of probabilities over the intervals. This treatment is referred to as *lmsr1*.

- In the third treatment the participants assessed the probabilities for each task and interval in a market featuring a LMSR. The modification regarding the second treatment is that the participants were not alone on this market. Therefore, the "final" prices of the last participant were shown to the new participant as the current market prices[5]. This treatment is referred to as *lmsr2*.

Lmsr1 and *lmsr2* together are referred to as *lmsr*. The differentiation between *lmsr1* and *lmsr2* allows to estimate the effect of anchoring in the market treatments at least qualitatively. The experiment design results in six groups of participants that were pseudo-randomly[6] assigned to one of the three treatments and then again pseudo-randomly to the high or low partition. Participants remained in their treatment in all three forecasting tasks. Participants were not aware of different treatments of intervals.

In line with Sonnemann et al. (2013), the participants were asked for an initial estimation and a final estimation of the absolute value before and after each forecasting task. After the participants finished all forecasting tasks, they continued with a "primary process test"[7] as a proxy for the used type of processing, which is based on Brakel, Shevrin, and Villa (2002). At the end a questionnaire on the basic demographics was performed that also included two self-formulated questions for each task. The questions asked (topic specific) if the participants have knowledge in this specific topic (have seen or informed themselves on the current value) and if they have experience in this specific topic (have a car or trade regularly)[8]. Both questions are binary coded.

Therefore, the experiment procedure is as follows:

1. Assignment to the treatments.
2. Instructions for the forecasting tasks (treatment specific).
3. Forecasting task (3x; one time for each of the three topics).
   - Initial estimation on the absolute value of the realization.
   - Expectation poll or market to assess the probabilities for each interval
   - Final estimation on the absolute value of the realization.
4. Primary process test.
5. Demographics questionnaire.
6. Debriefing.

Based on the considerations in Section 6.2, following three hypotheses were formulated:

- **$H_1$1:** The value for partition dependence is lower in the *lmsr* treatments compared to the *poll* treatment.
- **$H_1$2:** The portion of systematic processing measured by the primary process test is higher in the *lmsr* treatments compared to the *poll* treatment.

[5] In the unlikely event that two or more participants would have entered the experiment at the same time, participants would have seen each other trading. This, however, did not happen in current case.

[6] Participants were assigned to the treatments using round-robin according to the time of first entry into the experiment. Due to a higher rate of drop-outs in the market treatments, the poll treatment was closed and the rest of the participants were only distributed among the market groups.

[7] Implicit measurement of the current type of processing still very challenging (Samson and Voyer, 2012). Therefore, this experiment uses an explicit measures. The primary process test by Brakel, Shevrin, and Villa (2002), though a measure of a personality trait, seemed promising.

[8] See Table A.3 for the complete list of items.

TABLE 6.1: Three forecasting tasks of experiment 1 with partitions.

Question	Partition				
What will the price of one liter Diesel (Aral, Durlacher Allee) be at 12:00 am on Tuesday 20th December 2016?	High	<1.075	1.075 - 1.09	>1.09	€/l
	Low	<1.06	1.06 - 1.075	>1.075	€/l
What will the value of the "Deutsche Bank" stock be at the end of the trading day Tuesday 20th December 2016?	High	15.5	15.5 - 17	17	€
	Low	<14	14 - 15.5	15.5	€
What will the index of the DAX-30 be at the end of the trading day Tuesday 20th December 2016?	High	<10850	10850 - 11000	>11000	Points
	Low	<10700	10700 - 10850	>10850	Points

- **$H_1$3:** High experience and knowledge lead to a lower value of partition dependence.

Procedure

Students from a university group[9], whose primary purpose is to talk and exchange about stock markets, were invited to the survey and to forecast the realization of the three economic figures at a the 20th December 2016. Based on their interest, knowledge in the context of the three economic figures and trading in general can be assumed. Roughly 300 students were invited, 60 responded (participation rate = 20%), and 50 passed the selection criteria. Before the participants started with the forecasting tasks, they had to read the full instructions and answer two control questions to prove that they understood the task and, if applicable, how to trade reasonably. Participants were excluded in particular when they did not provide all answers or did not answer the control question correctly. The students were awarded the incentive of three amazon vouchers assigned according to the forecast performance. The chosen forecasting tasks and partitions for each question can be seen in Table 6.1.

In order to generate the intervals, one week before the experiment started, the current value of the economic figures to predict were chosen as a "midpoint". Based on this, the intervals were set by an educated guess of a student in the field of economics who regularly observed these prices. He was instructed to select the intervals in such a way, the probability for the two inner intervals should be approximately 50%. The dimensions of the intervals were not crucial for the experimental setting, as long as one interval would not receive almost 100% of the likelihood. Otherwise, it would not be possible to measure partition dependence. By an educated guess the width of the intervals should not be ill-sized in a dimension, where results should be affected significantly. The participants could enter the experiment from the 9th until the 19th December 2016.

The user interfaces are illustrated in Figure 6.4 and Figure 6.5. In the *poll* treatment, the intervals were presented as rows in a table in the first column. The second column included input forms to insert the estimated probability for a realization of

[9] "Börsen Initiative Karlsruhe e.V."

FIGURE 6.4: User interface of the *poll* treatment in the first experiment and the DAX-30 task.

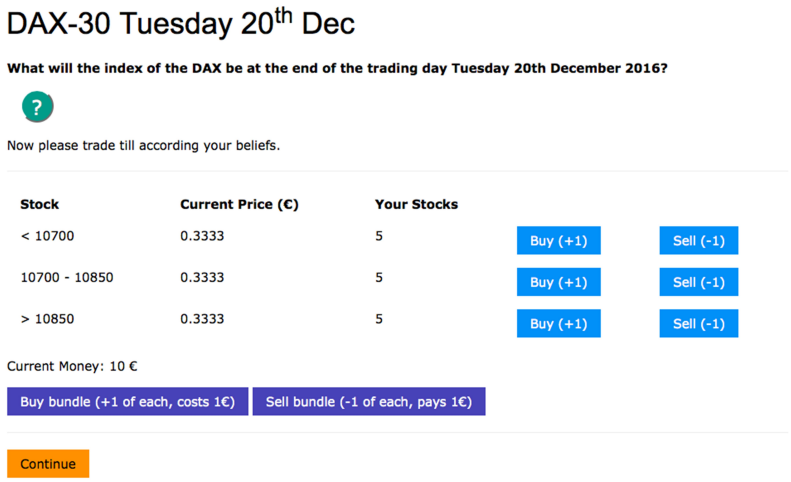


FIGURE 6.5: User interface of the *lmsr* treatments in the first experiment and the DAX-30 task. *Lmsr1* and *lmsr2* differ only in regard to the shown starting price (here: "current price").

the economic figure within this interval. In the *lmsr* treatments, the intervals were also presented in the first column. The second column presented the current price, and the third column the current stocks in the participant's depot. In a fourth column the handles to buy and sell were provided. In addition, the trade of "bundles" was possible with two more buttons below the table. A bundle is a package of one stock of each interval that could be sold or bought for 1 MU (as the likelihood for the realization of the economic figure in one of the disjunctive and together exhaustive intervals equals 100%). A text below the table indicated the money currently available to the user. Besides this the user interfaces looked similar.

After the forecasting tasks, the participants had to solve six "images" of the primary process test by Brakel, Shevrin, and Villa (2002), which were displayed one after another. Finally the participants had to complete the demographics questionnaire and had the opportunity opt-in for the Amazon voucher lottery. They also could opt-in to receive further background information on the experiment and its findings.

The length of the online experiment depended on the setting. It took between

roughly 4min and 30min (*poll*: in average 9.7 min, *lmsr1* in average 14.4 min, *lmsr2* in average 12.5, the difference between the *poll* and the *lmsr* treatments is significant, the difference between the *lmsr* treatments is not significant). In addition, the time it took to read the instructions was one of the main drivers for the length of the experiment.

6.3.2 Results

50 (female: 7, no answer: 3) participants completed the survey. Due to some drop-outs during the experiment, the distribution of participants over treatments is not completely equal (see Table 6.2). However, the distribution of gender, if applicable at all, does not show a significant difference (test for equal proportions, p=.613). The results for the partition dependence are reported in Table 6.2.

The overall forecasting error (deviation of average over all initial estimations and the realization) was 0.034 in the *poll* treatment and 0.014 in the *lmsr* treatments (3.4% and 1.4% of the absolute value). Therefore, it can be assumed that participants answered reliably. The *lmsr* treatments showed a significantly lower overall forecasting error (p=.031, Wilcoxon test, two-sided) than the *poll* treatment.

To test for $H_1 1$, the partition dependence in the three *treatments* is compared. A one-tailed[10] Wilcoxon rank sum test was applied to evaluate whether the partition dependence effect is significant. It has to be noted that current sample is missing statistical power for the Wilcoxon rank sum test to robustly find significance due to its small size and the presented reported p-values should be understood as trends. A comparison of the partition dependence bias between the *poll* and the *lmsr2* treatment showed that the partition dependence was less in the *lmsr2* treatment (ca. 0.01, no partition dependence) compared to the *poll* treatment (ca. 0.15, weak partition dependence). This finding supports hypothesis $H_1 1$. Strong partition dependence also occurred in the *lmsr1* setting (ca. 0.23, strong partition dependence). This finding would reject hypothesis $H_1 1$ and is an unexpected outcome when contrasted to *lmsr2*. This two contradicting results can be explained, when other disturbing factors are considered: It is presumed that this effect occurred due to the anchoring and adjustment bias (Campbell, Sharpe, and Others, 2009). In the *lmsr1* treatment, the participants were shown an initial pricing of 33ct for each stock. This priming may have been overloading the effect of the prior ignorance of partition dependence and, subsequently, have led to the strong occurrence of the shift in the estimations. Sonnemann et al. (2013) already warned of the occurrence of this effect. As in *lmsr2* the participants faced the estimation of their predecessor, this effect did not take place.

To test for $H_1 2$, the results of the primary process test are compared over the treatments. Based on the primary process test by Brakel, Shevrin, and Villa (2002), a "score for systematic processing" (SP score) (= 1 - "score for heuristic processing")[12] could be calculated. According to this SP score, most participants were classified into the systematic processing category. Participants seemed to choose pictures consistently of either the heuristic or the systematic category. To test for $H_1 2$ a simple linear regression was calculated to predict the SP score based on the treatment. A not significant regression equation was found ($F(2,47)=0.82$, p=.447), with R^2 of 0.034. A test for unequal distribution of heuristic and systematic processing also resulted in

[10] A one-tailed test was applied, as the theory behind partition dependence already induces the direction of the effect.

[11] Significance codes for all analyses: 0.000 *** 0.005 ** 0.01 * 0.05 . 0.1 n.s. 1

[12] "Rational" and "attributional thinking" in the wording of Brakel, Shevrin, and Villa (2002).

TABLE 6.2: Values for partition dependence in experiment 1. Significance[11] of the partition dependence values is tested with a one-tailed Wilcoxon test (sum of assessed probabilities of two neighboring intervals is greater than the probability on the corresponding interval in the other partition).

	N	Diesel price	Deutsche Bank stock	DAX-30
poll	21	0.12n.s.	0.18*	0.17.
lmsr1	15	0.25**	0.29***	0.14***
lmsr2	13	0.06.	-0.05n.s.	0.02n.s.
overall	50	0.14*	0.11*	0.15*

no significant unequal distribution (p=.213) between the *poll* and *lmsr* treatments. Therefore, for now, $H_1 2$ has to be rejected.

To test for $H_1 3$, the self-reported scores for experience and knowledge are compared over the treatments and tasks. Self-reported experience and knowledge for the DAX-30 task was approximately double of the experience and knowledge reported for the other tasks. The values for partition dependence for all treatments are continuously lower in the DAX-30 task, which already suggests that $H_1 3$ may be confirmed. To evaluate this hypothesis, the participants for each task are split into two groups according to their answers for experience and knowledge. In the expert group participants reported both to be true, in the non-expert group participants reported at least one to be false. In the DAX-30 tasks, experts occurred a weaker partition dependence bias (0.10.) than the non-experts (0.19*). In the Diesel tasks, experts occurred a stronger partition dependence bias (0.22**) than the non-experts (0.11n.s.). In the Deutsche Bank tasks, experts occurred a weaker partition dependence bias (0.14n.s.) than the non-experts (0.14.). Therefore, the results indicate towards a confirmation of $H_1 3$: Experts occur less partition dependence. However, due to several limitations (small sample size, not equally sized groups of experts vs. non-experts, questionable questions to query knowledge and experience) the results should be understood as first promising insights. Therefore, $H_1 3$ will be investigated experiment 2 more in detail.

The results of experiment 1 should be considered under several limitations. First, the sample size is yet not large enough to indicate more than trends. Second, it is necessary to discuss for which reason theory and results diverge regarding the reject of $H_1 2$. It may be possible that a second type error occurred. Another explanation, as already discussed above, is that the test of Brakel, Shevrin, and Villa (2002) actually is a measure for personality traits. Given these results, it may be possible that the treatments could not take enough influence to induce an effect that would be registered by this personality measure directly after the experiment. Therefore, it is to conclude that the primary process test of Brakel, Shevrin, and Villa (2002) seems not to be suitable to measure the effect of the treatments on the processing style at the exact moment. Third, the results of the *lmsr1* treatment may have been subject to a strong anchoring and adjustment bias. Finally, a high rate of drop-outs in the *lmsr* treatments was recorded, which may have led to a self-selection bias in the sample. However, still it is possible to derive some meaningful conclusions and amendments for the second experiment.

6.4 Experiment 2

Based on the results of the preliminary experiment 1 and feedback that was received from other researchers, the design of the first online experiment was reconsidered and a second experiment was performed. Besides small changes in the forecasting task, the focus is now also on further moderators of cognitive load (see moderators in brackets in the research model in Figure 6.3) that seemed to have had an influence in the first experiment. In addition, the sample of subjects was enlarged.

6.4.1 Method

Stimuli & Design

The design of the experiment consisted of two forecasting tasks. Following the paradigm of experiment 1, in the first task the state space is partitioned two times into three intervals, shifted by one interval (see Figure 6.2). The second forecasting task did not implement a shift of the state space, but rather an extension of the state space and is more comparable to the "fault-tree" setting in Experiment 5 of Fischhoff, Slovic, and Lichtenstein (1978) (see Section 6.2 for further details). Therefore, in the second task, the state space is partitioned one time into three intervals and one time into four intervals. Here, the last two intervals of the more granular partition equals the last interval of the less granular partition. In each task, all participants had to estimate the probability of the outcome of an event (continuous variable in the first forecasting task, discrete variable in the second forecasting task) falling into the predefined intervals.

As in the first experiment, a *poll* treatment was conducted. In contrast to the first experiment, only one *lmsr* treatment was conducted. The *lmsr1* treatment from experiment 1 was dropped, as it was not possible to distinguish between anchoring and adjustment effect from the initial prices and effects induced by the partition dependence bias. In addition, the *lmsr2* treatment from experiment 1 was modified, resulting in a new *lmsr3* treatment. In the new *lmsr3* treatment, the participants traded on a market against an "artificial" trader and not against "human" traders. The initial prices were set to the uniform distribution (ignorance prior). After five seconds, an artificial trader started trading. The underlying preferences of this trading equaled to the last inputs of a participant in the *poll* treatment. The artificial trader traded the assets (intervals) in a random order and at random points of time (bounded to 25 seconds at the latest). The introduction of other traders in the market in the form of an artificial trader allowed the human participant to realize "profits", in case his estimation differs from the estimation of the artificial trader. The human participant was informed in the instructions that there may be other traders in the market. As the estimation of the artificial trader was taken from the last poll estimation, it was ensured to be "reasonable" and "realistic" to some extent and, at the same time, random. Thereby, it was ensured that the estimations in the *lmsr3* setting were more independent of each other[13]. This is important, as median splits on the data based on diverse criteria will be performed for the evaluation. Participants were randomly assigned to either the *poll* or the *lmsr3* treatment.

Similar to Sonnemann et al. (2013), participants were asked for an initial estimation of the absolute value before the forecasting task. The final estimation after

[13] In the lmsr2 treatment the estimation of a trader was influenced by the estimation of the previous trader. This effect is reduced in the lmsr3 setting, as the first numbers shown to the trader always equals the uniform distribution.

each forecasting task was dropped, as participants in the first experiment reported this as confusing. Before the forecasting tasks, participants answered a German version[14] of the *Rational-Experience-Inventory Short-Version (10)* (REI-10[15])[16] from Epstein et al. (1996) and solved a conjunction fallacy problem from Tversky and Kahneman (1983) ("Risky Choice"). It was intended to test their individual preference for either heuristic or systematic processing. After every forecasting task, the participants answered a questionnaire adopted from the questionnaire introduced by Trumbo (2002), which measures the current (topic-specific) processing style. This questionnaire was adapted to each topic and translated into German[17]. Subsequently, a questionnaire regarding their topic-specific *expertise* was conducted. The items for the construct *expertise* were self-formulated and as follows (translated from German)[18]:

- Could you tell someone a lot about this topic?
- Would you read a newspaper article on this topic?
- Do you think, a friend would ask you for help or advice regarding this topic?
- Did this topic already affect you?

Items were assessed on a seven point Likert-scale. Finally, participants received a questionnaire regarding basic demographics and some further questions on *experience with markets* (self-formulated item), *motivation* (self-formulated item), and *risk* (Dohmen et al., 2011).

Therefore, the experiment procedure is as follows:

1. Assignment to the treatments.
2. Rational-Experience-Inventory Short-Version (REI-10).
3. Conjunction Fallacy Problem ("Risky Choice").
4. Instructions for the forecasting tasks (treatment specific).
5. Forecasting task (2x; one time for each of the two topics).
    - Initial estimation on the absolute value of the realization.
    - Expectation poll or market to assess the probabilities for each interval.
    - Questionnaire to measure the topic-specific information processing style (adapted from Trumbo (2002)).
    - Questionnaire to measure the topic-specific expertise.
6. Demographics questionnaire (including questions on motivation and general experience with markets, etc.).
7. Debriefing.

Based on the considerations in Section 6.2 and the results of the first experiment (Section 6.3), the following hypotheses were formulated:

- **$H_2$1:** The value for partition dependence is lower in the *lmsr3* treatment compared to the *poll* treatment.
- **$H_2$2:** The portion of systematic processing is higher in the *lmsr3* treatment compared to the *poll* treatment.
- **$H_2$3:** High *experience with markets* leads to a lower value of partition dependence.

[14] own translation

[15] The REI-10 measures the two reflective constructs "need for cognition" and "faith in intuition" by five items on a 5 point Likert-scale from completely false to completely true. The five items were selected from a longer version of the test (REI-42), selected by the highest factor loading each.

[16] See Table A.4 for the complete list of items.

[17] See Table A.5 for the complete list of items.

[18] See Table A.6 for the complete list of items.

TABLE 6.3: Two forecasting tasks of experiment 2 with partitions. (Union = U, SPD = S, Green Party = G, FDP = F)

Question	Partition				
What will be the price of the next iPhone at its introduction? (translated from German, in €)	High	<950	950 - 1000	>1000	
	Low	<900	900 - 950	>950	
What will be the ruling coalition after the next Federal Election in Germany? (translated from German)	High	U-S	U-G	U-F-G	Others
	Low	U-S	U-G		Others

- **$H_2$4:** High *motivation* leads to a lower value of partition dependence.
- **$H_2$5:** High *expertise* leads to a lower value of partition dependence.
- **$H_2$6:** A high "systematic processing style" (low heuristic processing style) as a personality trait leads to a lower (higher) value of partition dependence.

Procedure

The participants were asked to forecast the price of the next iPhone (2017) and the members of the future ruling coalition[19] after the next German Federal Election. These topics were chosen according to the consideration that they might allow a "relatively" clear distinguishing between knowledgeable participants (e.g., "Apple user") and not knowledgeable participants (e.g., "Android user"). The chosen forecasting tasks and partitions for each question are reported in Table 6.3.

The intervals/options were chosen by the educated guess of the experimenter under considerations of the likelihood of the events. Therefore, the width of the intervals should not be ill-sized in a dimension, where results should be affected significantly. The task for the experiment was opened at the 21st June 2017 and closed at the 7th July 2017. The experiment was conducted in the context of a research seminar. Two students spread the invitation to the experiment that was hosted on local servers. Participants were given the incentive of the chance to win one of ten vouchers for a local ice-cream shop, popular among students, worth €55 in total. The reach of the invitation could not be measured reliably, however, ca. 700 participants opened the link in the invitation from which 109 person responded and 80 passed the selection criteria. Participants were excluded when they did not provide all answers, did not answer the control question correctly, or provided no or non-sense values for most likely coalition, or the most likely price of the iPhone (lower €500 and larger €1500) during the initial estimation (free text answers), before the forecasting task. Before the participants started with the forecasting tasks, they had to read the full instructions and answer two control questions to prove that they understood the task and, if applicable, how to trade reasonably.

In contrast to the first experiment the user interfaces and instructions for the experiment were modified as follows:

[19] In the German election system usually no party is able to receive more than 50% of the votes. Hence, several parties have to form coalitions in order to gain a majority in the parliament to be able to govern.

FIGURE 6.6: User interface of the *poll* treatment in the second experiment and the iPhone task (in German).

- A grey placeholder text was added into the fields in the *poll* treatment of "1/number of options" (e.g., 33.3%), to ensure that both treatments face the same number at the first moment (see Figure 6.6).
- A small text at the bottom of the form was added in the *poll* treatment that sums up the numbers provided in the fields, to ease the adding up (see Figure 6.6).
- The option to buy and sell bundles was dropped, as this option was only rarely used in the first experiment. In addition, participants responded sometimes that they did not quite understand the utility of these buttons. Finally, the explanation of this functionality required a lot of space in the instructions, while it provided only little added value.
- Participants were told and, therefore, aware of the fact that other participants may see other intervals as suggested by Sonnemann et al. (2013) to lower the effect of anchoring and adjustment.

The length of the online experiment depended on the setting. It took between roughly 6min and 30min (*poll*: in average 12.2 min, *lmsr3* in average 13.7 min, difference is not significant), potentially depending on how quick participants understood the market.

6.4.2 Results

80 (female: 27) participants completed the survey. Due to some drop-outs, the number of participants over treatments is not completely equal (see Table 6.4). The gender is equally distributed over the treatments[20].

Due to the limited number of participants, the threshold for acceptable reliability of the constructs was set to a Cronbach's α of 0.7. For the *REI-10* Cronbach's α was 0.86 ("faith in intuition" = heuristic processing) and 0.74 ("need for cognition" = systematic processing). For the questionnaire of Trumbo (2002), Cronbach's α was 0.79, 0.46, 0.82, and 0.74 (heuristic processing in iPhone tasks, heuristic processing in coalitions task, systematic processing in iPhone task, systematic processing in coalitions task). Besides the construct "heuristic processing" in the coalitions task, that did not exceed the threshold of 0.7, the results of these questionnaires can be used in the evaluation. As all these constructs were pre-validated by other researchers it is not clear why the necessary reliability could not be reached in one case. This case will, therefore, be treated with caution in the discussion.

[20] A two-sample test for "equalitiy of proportions" (prop.test()) shows no significant inequality (p=.250) regarding the gender.

TABLE 6.4: Values for partition dependence in experiment 2. Significance of the partition dependence values is tested with a one-tailed Wilcoxon test.

	N	iPhone	coalitions
poll	45	0.16*	0.07n.s.
lmsr3	35	0.14n.s.	0.08n.s.
overall	80	0.15*	0.08*

The construct for *expertise* with self-formulated items received good reliability. In the case of the iPhone task, the reliability of this construct was excellent (Cronbach's α=.92). In the case of the coalitions task, the reliability was still good (Cronbach's α=.85).

The values for partition dependence are reported in Table 6.4.

Both tasks showed a significant partition dependence bias that dissolves, when the data is split according to the treatments. In the case of the iPhone task, the significance preserves for the *poll* treatment. In the case of the coalitions task, the significance dissolves in both treatments. The susceptibility of the significance indicator to the split, indicates that for this kind of evaluation more data would be profitable. Therefore, hypothesis $H_2$1 would be confirmed in the iPhone task, rejected for the coalitions task. This discrepancy at this point is a strong indicator that the effect of the partition dependence bias is not only moderated by complexity, but other moderators that are addressed by the hypotheses $H_2$2-6.

In the coalition task, the measures for heuristic and systematic processing by Trumbo (2002) between the *poll* and *lmsr3* treatment did not differ significantly. In the iPhone task, the measures for systematic processing by Trumbo (2002) between the *poll* and *lmsr3* treatment did not differ significantly. Heuristic processing was significantly higher (p=.055). However, these results has to be treated with caution due to the low Cronbach's alpha (0.46). Nevertheless, these findings are consistent with the partition dependence values that only differed in the iPhone task. Therefore, both indicates towards a confirmation of hypothesis $H_2$2. Analogously to experiment 1, the limitation has to be kept in mind that the measures of Trumbo (2002) do measure the topic-specific degree of heuristic and systematic processing and not necessarily the current state.

To test for the hypotheses $H_2$3-6, median splits on the overall data, not considering the treatments, are performed. However, it is necessary to discuss the limitations of this procedure in advance in order to understand the results and their implications. Not considering the treatment would be legitimate if it can be assumed that the treatment has no influence on the estimation. As shown in Table 6.4, this is only the case for the coalitions task, but not the case for the iPhone task. Therefore, this assumption cannot be finally verified. However, due to the fact that partition dependence can only be measured by aggregating the estimations of a group, the control for a treatment effect by using multivariate statistics (e.g. multiple linear regressions) also disqualifies. This limits the implications derived of the data of the iPhone task.

To test for $H_2$3, a median split on *experience with markets*, a 7 point Likert-scale item collected within the demographics questionnaire (What is your experience with markets: "low" to "high"), was performed. The participants were split by >4 (high) and <=4 (low). Participants with high *experience with markets* showed no significant partition dependence bias in both tasks. Participants with low *experience with markets* showed weak significant partition dependence bias in both tasks (coalitions: p=.053;

iPhone: p=.079). The reasoning behind this hypothesis argues that participants that do not quite understand or do not feel comfortable with a market setting are more likely to rely on heuristics. This hypothesis seems to be confirmed.

To test for $H_2$4, a median split on *motivation*, a 7 point Likert-scale item collected within the demographics questionnaire (How did you perceive the survey: "bothersome" to "very interesting"), was performed. The participants were split by >4 (high) and <=4 (low). In the case of the coalitions task, a significant partition dependence is found in the group with high *motivation* (p=.029), while no partition dependence is found in the group with low *motivation*. In the case of the iPhone task, the opposite pattern is found: No partition dependence in the group with high *motivation* and significant partition dependence in the group with low *motivation* (p=.005). Therefore, the hypothesis is rejected.

To test for $H_2$5, a median split on the construct *expertise* was performed. The participants were split by >4 (high) and <=4 (low). Participants with high *expertise* showed no significant partition dependence bias in both tasks. Participants with low *expertise* showed a (weak) significant partition dependence bias in both tasks (coalitions: p=.015; iPhone: p=.088). These findings confirm hypothesis $H_2$5.

The results of the REI-10 showed, as expected, no significant differences between the treatments. To test for $H_2$6, a median split (>=4 heuristic processing, >=4.125 systematic processing) on the REI-10 measures was performed. Participants with high heuristic processing showed significant partition dependence in both tasks (iPhone: 0.19*, coalitions: 0.1.). Participants with low heuristic processing showed in both tasks no significant partition dependence (iPhone: 0.10n.s., coalitions: 0.03n.s.). Accordingly, the partition dependence values for a median split for systematic processing turn to be significant for the group with low systematic processing (iPhone: .16*, coalitions: .12*) and insignificant for the group with high systematic processing (iPhone: 0.12n.s., coalitions: 0.00n.s.). These results indicate that personality had a much greater influence on the applied processing style of each individual participant than the different complexity of the treatments. Nevertheless, these findings confirm hypothesis $H_2$6.

6.5 Discussion

The results of each experiment have been directly discussed in the corresponding subsections of Sections 6.3 and 6.4. Persistently, in both consecutive experiments the occurring of the partition dependence bias could be observed in the forecasting tasks. The method of Sonnemann et al. (2013) to measure partition dependence proved expedient. The new method to add an option, as suggested in the second experiment, which was more leaned at the "fault trees" setting, showed less partition dependence. Yet, it cannot be estimated if this is due to the difference between numerical intervals and discrete options, more expertise in this topic in general (Fox and Clemen, 2005), or other factors. The new method is, therefore, suggested for evaluation in further experiments.

Both experiments showed that complexity, as a suggested moderator of extrinsic cognitive load, seems to moderate the occurrence of partition dependence according to the Heuristic Systematic Model. The two experiments suggest that this hypothesis is true, though the measurement and confirmation proved to be difficult. A definitive confirmation would need an elaboration with more statistical power and a more reliable measure for the current processing style. The presented experimental evaluations especially lacked of an implicit or valid explicit measure of current

cognitive load (or type of processing) at the moment of the experiment. Therefore, a laboratory experiment may be considered in future research, which allows to apply physiological measures. Suggested measures are, e.g., the measurement of pupil dilation (Brünken, Plass, and Leutner, 2003). Jung and Dorner (2018) are applying the measurement of arousal to distinguish between the processing styles in the context of decision inertia, another cognitive bias in decision-making.

The results of the first experiment indicated that complexity was not capable to fully explain the occurrence of partition dependence, for which reason the second experiment design controlled for several more moderators. The resulting data supports the moderating effect of expertise: High knowledgeable individuals are less susceptible to partition dependence. The results also support the hypothesis that a high experience with markets leads to less partition dependence. This effect can be attributed to cognitive depletion (Chen et al., 2017). Individuals without experience with markets do not quite understand the strategy to realize (financial) profits by trading according to an optimal strategy and fall back to the use of heuristics (and heuristic processing). Motivation was also suggested as a moderator of partition dependence (high motivation leads to low partition dependence) (e.g. Watts, 2015). This could not be confirmed. However, it cannot be assured that the two treatments provided sufficient diversity on the range of motivation. In the presented experiment, motivation was measured by a single item self-assessment scale. In future research this should be further broken down in order to differ between motivation induced by the treatment and intrinsic motivation induced by the topic. Heuristic and systematic processing as a personality trait proved to be a strong moderator to the occurrence of partition dependence, potentially stronger than the processing style induced by the implemented treatments. One can derive two conclusions from this: First, the treatments should be designed even more different regarding "complexity" in order to increase the influence on the processing style (e.g., using a CDA). This, however, comes with the risk of triggering cognitive depletion. The strong occurrence of partition dependence in Sonnemann et al. (2011) and Sonnemann et al. (2013) can be regarded as a hint towards this conclusion. Second, the complexity of the task should only be regarded as one of many moderators on the processing style and always be examined within the context of other moderators. With this evidence, it may be suggested to detach the complexity from the market mechanism towards the complexity induced by the "expression of probability estimations in form of (real) money" (Levin, Chapman, and Johnson, 1988).

Finally, the Heuristic Systematic Model (as one representative of the Dual-Process Theories) proved to be a valid framework to explain the occurrence of the partition dependence bias partly in forecasting tasks. Though it is not possible right now to rank the moderators according to their share on the occurrence of partition dependence, expertise and the personality trait towards a systematic processing style seem to explain a significant portion of the effects. Such a ranking might be achieved by future work and a dedicated experiment design towards this question. Nevertheless, the presented experiments can be regarded as an exploratory study on the applicability of the Dual-Process Theories to explain this phenomenon. Hereby, this chapter shed first light on future pathways in this research field. There remains the need of addressing each individual moderator more in detail. There remains also the need to evaluate other cognitive biases in managerial decision-making, besides partition dependence, from the Heuristic Systematic Model perspective.

Some limitation have already been addressed within the results of each experiment. First, the modest sample sizes in both experiments put a strong limitation to the measurement of the strength of the effects. Especially as the measurement of

partition dependence is not possible on an individual's level, a larger sample would probably make the effects more consistent. Second, Brakel, Shevrin, and Villa (2002) showed that heuristic and systematic processing as a personality trait are subject to change during a lifetime. All participants in both experiments were students. Therefore, it is not possible to exclude that some moderators do have different impact in other samples (e.g., managers). Third, both attempts to measure the current processing style (primary process test by Brakel, Shevrin, and Villa (2002) and the self-reporting questionnaire by Trumbo (2002)), do not seem applicable to measure the current processing style beyond doubt. Finally, the REI-10 and the measures by Trumbo (2002) were translated into German in the second experiment to ensure that the participants understood each question. It may be necessary to re-validate these questionnaires in case that they will be used in the future again.

6.6 Conclusion on Cognitive Factors

In this chapter the occurrence of partition dependence in the context of forecasting was demonstrated and additional evidence to the existence of this phenomenon is contributed. The necessity to carry out more research to understand the occurrence, as well as driving and moderating factors was underpinned. The presented experiments and some other academic work demonstrated the persistence of partition dependence in decision-making. However, only little research was done regarding the negative effects of partition dependence on the outcomes of biased investment decisions and forecasting, though the problem is undeniable (Bardolet, Fox, and Lovallo, 2011). This study utilized the framework of the Heuristic Systematic Model to explain and understand the occurrence of partition dependence. The results of the two consecutive experiments demonstrated its applicability. It also contributed to the current research on partition dependence by reproducing the findings of other authors that expertise decreases the influence of partition dependence (e.g. Fox and Clemen, 2005). Though it was not finally possible to demonstrate that complexity induced by the mechanism of estimation submission influences the type of processing and, thereby, partition dependence, the chapter still demonstrated the effect of several other moderators (e.g., expertise, "need for cognition", and experience with markets). Especially an individual personality trait towards systematic processing (or heuristic processing) lowers (raises) the probability to be susceptible to partition dependence.

Here, a focus was laid on partition dependence, which is, however, not the only cognitive bias that may be explained by approaching findings from the Dual-Process Theories. By applying the Heuristic Systematic Model to explain the partition dependence, this chapter also contributes first conceptual work towards a deeper understanding of the cognitive biases in group decision-making and forecasting and in the context of IS from a Dual-Process Theories' perspective. The presented experiments showed that there is not one main driver for the processing style. The effects and moderators mentioned in the current study need to be evaluated in future research more in detail to provide resilient and quantifiable evidence. There probably won't be a serum against cognitive biases in the near or far future, therefore, decision makers, forecasters, prediction market traders, and so on should be advised to be aware of potential irrational influences, such as the partition dependence bias. Using markets for expectation elicitation or training decision makers regarding individual expertise and personality may be some options to reduce its effect. However, they do not moderate its existence completely.

Chapter 7

Improving Response Quality: Motivational Factors

> You shall not steal; you shall not deal falsely; you shall not lie to one another
>
> Leviticus 19,11; ESV

7.1 Problem Formulation

Contents of this section are in part adopted from Kloker et al. (N.D.).
See Section A.1 for further details.

In Section 2.2 the wide adoption of prediction markets in political and organizational decision making was already emphasized. Companies like HP, Deutsche Telekom, Google, Yahoo, General Electrics, and numerous others use prediction markets to support their idea and innovation management, idea generation and project monitoring (Bothos, Apostolou, and Mentzas, 2009; Cowgill, Wolfers, and Zitzewitz, 2009; Graefe, Luckner, and Weinhardt, 2010; LaComb, Barnett, and Pan, 2007; Mangold et al., 2005; Soukhoroukova, Spann, and Skiera, 2012; Spears and LaComb, 2009; Rohrbeck, Thom, and Arnold, 2015). In many of these cases large financial resources are allocated, reputation and sometimes even jobs are tied to these decisions. However, when the stake is high, many stakeholders are willing to exert influence wherever possible which leads to manipulative and fraudulent attempts on the market. This also applies to prediction markets, which have a high potential to influence decisions and even public opinion through their forecasts (Rhode and Strumpf, 2008). The most advanced prediction markets have a fraud (and manipulation) detection software component that runs continuously on the market, scans every interaction and transaction, and looks for fraudulent patterns (Kloker and Kranz, 2017; Schröder, 2009; George Mason University, 2015). Especially, but not only, in play-money prediction markets patterns that wear out one account in favor of one or more other accounts are observed (Blume, Luckner, and Weinhardt, 2010). The aforementioned algorithms detect such patterns, but

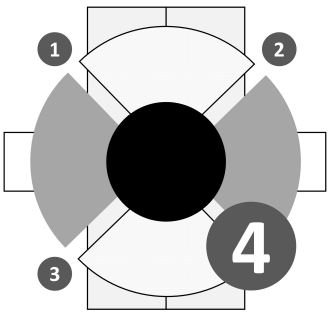


FIGURE 7.1: The presented research project in this section addresses the fourth source of errors according to the JFIM: Motivational biases.

are limited to the following aspects: They only find (1) known fraudulent patterns and (2) cases of suspected violations of the rules of the prediction market.

The first limitation, the problem of limited creativity in fraudulent pattern recognition of rule based algorithms is a problem that is regularly observed during the operation of a prediction market. Fraud is a phenomenon that especially (but not only) occurs in play-money prediction markets. In the FAZ.NET-Orakel, e.g., several cases were observed, where a single participant created multiple accounts and started to transfer money from one to another account by applying different trading patterns. Some of these fraudulent trading patterns are described in literature (Blume, Luckner, and Weinhardt, 2010) and are usually already implemented in the detection algorithms of current prediction markets. Anyway, a not known number of further patterns is not and prediction market operators are regularly surprised by the creativity of fraudulent participants.

This, however, is not the complete problem. The second limitation is even more problematic and deals with trading behavior, which is obviously manipulative on the one hand, but not punishable or demonstrable on the other. On a regular basis, literature or anecdotal evidence from the FAZ.NET-Orakel provide reports of cases, where a trader trades contracts of specific, often left- or right-wing political parties as if their expectation for this party would be a vote-share of, e.g., 85%. As these parties usually receive around 5%, reasonably it is to assume that the trader raised the price of this contract for other purposes than behaving as a "homo economicus", e.g., promoting the party[1]. However, right now a prediction market provider has no remedies against this obviously manipulative behavior that potentially harms the accuracy of the forecast, as he cannot accuse participants for having certain opinions. The trader can claim this in the event of a countermeasure by the market provider and knit accusations of suppressing freedom of opinion. Such cases can be observed in literature, but also in the current prediction market FAZ.NET-Orakel (Kranz et al., 2014; Hansen, Schmidt, and Strobel, 2004; George Mason University, 2015).

The peril of manipulation is one of "the five open questions about prediction markets" (Wolfers and Zitzewitz, 2006a) and "[...] is recognised [!sic] as a potential stumbling block to wide scale practitioner adaption" (Buckley and Doyle, 2017, p. 612). Problems from other contexts, with comparable properties are addressed in literature, e.g., regarding fraud in elections or in crowd-work tasks (Hirth, Hoßfeld, and Tran-Gia, 2011; Bader, 2013). These previous works suggest and utilize to crowd to detect and address manipulative and fraudulent issues. As prediction markets are a crowd-based tool as well, a crowd-based approach for manipulation and fraud detection seems promising and feasible. However, how to design and operate such a tool is an open question. In addition, there is no quantitative evaluation nor qualitative experience of the general applicability in existence.

To tackle these problems, an IT artifact was implemented in the prediction market FAZ.NET-Orakel and, therefore, subject to several constraints regarding the design, functionality, communication, and participants. To do justice to this specific context, the ADR methodology by Sein et al. (2011) was applied. The ADR methodology addresses two challenges (Sein et al., 2011, p. 40): (1) addressing a problem situation encountered in a specific organizational setting by intervening and evaluating; and (2) constructing and evaluating an IT artifact that addresses the class of problems typified by the encountered situation. In addition, designing *ensemble artifacts*, according to ADR involves dimensions beyond the technological, as they are also the result of design efforts regarding the contextual factors throughout the

[1] See Section 7.2.1 for a discussion on other reasons.

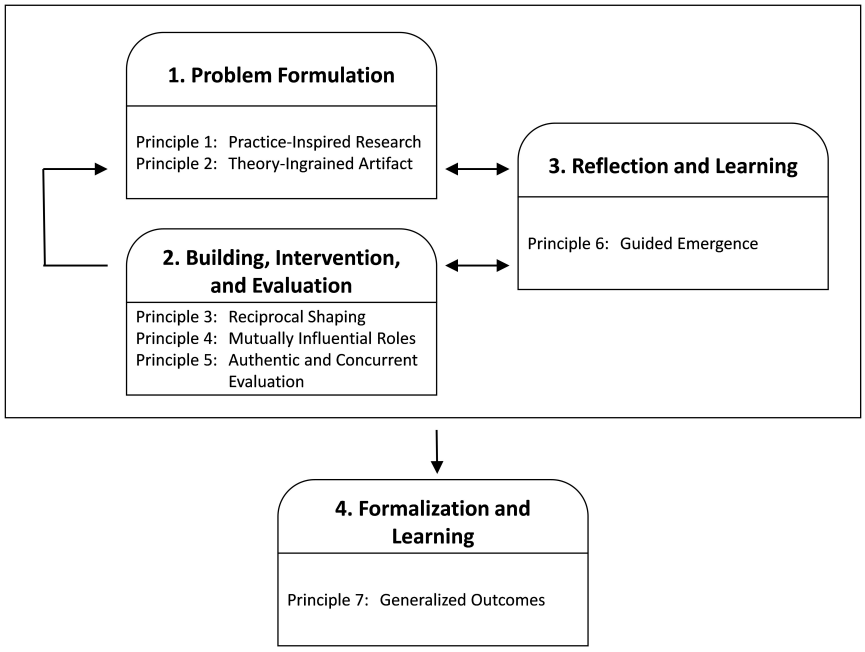


FIGURE 7.2: ADR Method: Stages and Principles. Adapted from Sein et al. (2011, p. 41).

design process. This meets both, the restrictions by the industry partner, and the specific problem. As manipulative and fraudulent actions cannot be induced intentionally within this context, controlled evaluation efforts are difficult to design and conduct. ADR helps to address this important issue adequately at this point. Figure 7.2 illustrates the ADR method according to Sein et al. (2011). The phase of evaluation is not divided from the implementation, as the implementation is already strongly driven by non-neutral interventions of the context of the subsequent evaluation. The phase of reflection and learning (3) is an ongoing process during the problem formulation (1) and implementation (2).

ADR emphasizes not to address problems merely as software engineers, who develop a program, nor merely as organizational consultants, who would only address the social level, but the intersections and interactions (Sein et al., 2011). Therefore, the current work does not only suggest a tool for detecting manipulation and fraud, but also considers what is necessary to run such a tool successfully. The central problem is classified into the class of *monitoring in the context of ambiguous and rare events*.

In the sense of Sein et al. (2011) the problem is addressed of a socio-technical perspective and three distinct, but interwoven research questions, are formulated that should be addressed by the artifact:

- (I) How to enable the crowd to detect creative manipulative and fraudulent patterns in contexts where (training) data is sparse and events cannot be induced intentionally?
- (II) How to decide on ambiguous manipulative and fraudulent cases that are not obviously against the rules.
- (III) How to ensure a continuous and reliable operation of such an artifact?

The objective of the current research is to define DPs for this class of problems and illustrate design considerations and implications, as well as consequences and experiences from the designed artifact. The remainder of this chapter is structured as follows: Section 7.2 references the current state-of-the-art in academic research in manipulation and fraud in prediction markets and, thereof, derives the detailed problem (problem formulation). In addition, it identifies and explains underlying kernel theories and refers to approaches to similar problems in other contexts. Section 7.3 then explains the considerations behind the suggested approach, as well as the context (project partners and organization) and the course of events. Section 7.4 describes the Building, Intervention, and Evaluation (BIE) phase of the ADR project and the IT artifact in detail. Section 7.5 then derives DPs based on the findings, states limitations of the artifact development and evaluation, and draws up future research possibilities.

Section 7.6.1 then switches the focus from detection to prevention and considers the role of motivation and incentive engineering and puts a fourth research question:

- (IV) How to design prediction markets in order to prevent manipulation and fraud?

Section 7.6.4 briefly introduces and summarizes the design adaption derived from previous considerations on the FAZ.NET-Orakel to prevent manipulation and fraud.

7.2 Related Work

> Contents of this section are in part adopted from Kloker and Kranz (2017) and Kloker et al. (N.D.).
> See Section A.1 for further details.

In the first part of this section, Principle 1 of the ADR methodology is fulfilled, *Practice-Inspired Research*, as it starts by reviewing the current academic discussion regarding manipulation (and fraud) in prediction markets[2]. First, the discussion regarding short- or long-lived effects of manipulation on forecasting accuracy is emphasized and summarized, where it is concluded that long-lived effects on the forecasting performance are not common but possible and, therefore, may cause problems for the innovation processes in companies, political judgment, or organizational decision-making in some cases where the stakes are high. Thereafter, cases of manipulation and fraud are summarized in order to give the reader an impression on the dimensions of this phenomenon. Third, the cases in the FAZ.NET-Orakel are described. Finally, current fraud detection strategies are briefly discussed and a need for other approaches is derived, which results in the problem formulation.

In the second and third part of the current section, Principle 2 of ADR is fulfilled, *Theory-Ingrained Artifact*, as generalizable theories are identified from previous work in IS regarding crowd-sourcing, decomposition of complex problems and creativity, especially Type V theories (Design and Action) (Gregor, 2006). Where necessary, the search for kernel theories also falls back on kernel theories from social psychology regarding motivation. In addition, literature on manipulation and fraud detection in other contexts is summarized.

[2]Literature that solely focuses on manipulation of the outcome as observed by Ottaviani and Sørensen (2007) and described in addition by Kloker and Kranz (2017) is explicitly excluded, as this is not fraud within the bounds of prediction markets, though it may be motivated by them.

7.2.1 The Problem of Manipulation and Fraud in Prediction Markets

Distinguishing Fraud and Manipulation

In line with Schröder (2009) this dissertation distinguishes manipulation from fraud in prediction markets: Manipulation is a "[...] speculative attack that achieves its objective of changing prices" (Rhode and Strumpf, 2008) *and* at that, playing according to the rules[3]. Fraud is a speculative attack by using measures that are not according to the rules. The objective of fraud may comprise several more objectives than changing prices.

Manipulation is an inherent challenge for prediction markets for several reasons. "Investors" can distort the predictions (prices) away from the fundamentals, e.g., only in order to influence the expectations and actions of others (Rhode and Strumpf, 2008). According to Rhode and Strumpf (2008, p. 1f), "[...] the potential reward from a successful manipulation can far exceed the financial resources needed to implement it." Let us consider the case of PSMs[4]: The current prices are often regarded as the true odds and therefore can influence the decisions of voters that are unwilling to support a politician that is not likely to have a real chance. Changes in the prices that make the politician appear to be in a competitive position, according to the predictions, "[...] require relatively small stakes [...] but can shape a very large outcome [...]" Rhode and Strumpf (2008, p. 1f). To put it in other words, in PSMs, manipulators want to influence the choice of undecided voters (Rhode and Strumpf, 2008; Hansen, Schmidt, and Strobel, 2004). Higher visibility raises the likelihood to be subject to manipulation (Rhode and Strumpf, 2008). In addition, prediction markets often do not prohibit trading by insiders (Rhode and Strumpf, 2008) that are capable to generate large profits from distorted prices. However, there is consensus that in most prediction markets usually manipulation does not intend financial gains for the time being, but is interested in the feedback effect of the prices (Rhode and Strumpf, 2008).

Does Manipulation harm Accuracy?

Especially regarding the effect of manipulation on prediction market prices some research was done and is still ongoing, with a preliminary slight consent that manipulation has no long-lived effects (Buckley and Doyle, 2017) on the accuracy of the predictions, though this cannot be finally concluded.

Based on the observation of real-world data from the TradeSports prediction market, Rhode and Strumpf (2008) "[...] find little evidence that political stock markets can be systematically manipulated beyond short time periods." This, however, is the prevailing opinion in literature (among others shared by Wolfers and Zitzewitz (2004) and Camerer (1998)). Oprea et al. (2008) studied the accuracy of forecasts by a third party based on information from manipulated markets. They concluded that the forecasts of the third party were not affected, though even half of all traders were given an incentive to change the third parties forecast. Hanson, Oprea, and Porter (2006) showed in an experiment that manipulation has no long-lived effects on market prices, if non-manipulative traders suspect manipulators to be active and are aware of their incentive and the direction of their actions. Hanson and Oprea (2009)

[3] "A successful manipulation is usually not possible unless the trades influence the beliefs of other market participants (An investor's beliefs are defined with respect to the fundamentals, as well as the future actions and beliefs of other investors)." (Rhode and Strumpf, 2008, p. 6)

[4] Political Stock Market

developed a theoretical model of incentives and manipulation and even highlight the liquidity-providing effect of manipulation[5].

In contrast to this tenor, Deck, Lin, and Porter (2013) argued that previous findings were all based on market settings, where manipulators suffered financial losses associated with their trading. Consequently, Deck, Lin, and Porter (2013) demonstrated in the same setting as Oprea et al. (2008) that well-equipped manipulators with high incentives can distort market prices to a level that (third party) forecasters are misled to predictions that are no better than random guessing — provided that inexperienced traders exist in the market. In this setting, Deck, Lin, and Porter (2013) observed two further interesting findings: First, the manipulative traders often ended up with positive trading profits (which is quite counter-intuitive and concerning) and, second, primarily the bids were influenced, while the asks remained informative.

Recent studies on manipulation in prediction markets focus on markets featuring a market scoring rule[6]. In this context Buckley and Doyle (2017) conclude in line with literature in traditional market settings that manipulation has no lasting impact on market prices. They demonstrated these findings in a field-experiment with 67 undergraduate students and 16 manipulative accounts. An earlier study with markets based on market scoring rules was performed by Jian and Sami (2012). In a two-participant market setting it was demonstrated that manipulative strategies, such as "bluffing" or "delaying" lead to a slower convergence of the market price but no further differences. However, Chen et al. (2015) demonstrated a successful manipulation in a market featuring a LMSR. Their setting included, exogenous instead of endogenous[7] incentives and similar to Jian and Sami (2012) only two traders. This was already suggested by Chen et al. (2007) in theoretical considerations for dynamic parimutuel markets and markets featuring a LMSR.

Summing up, it can be concluded that prediction markets can be manipulated to (at least) some extent and that this is potentially dependent on several properties of the individual context, such as the incentives.

(Further) Cases of Manipulation

In this section some cases of manipulation and or fraud that are reported in literature are summarized in order to gain an understanding of relevant factors and facets. A further collection of the numerous cases is provided in Table 7.1.

Rothschild and Sethi (2016) found suggestive evidence in the 2012 U.S. Presidential Elections in favor of Romney (against Obama) on the prediction market Intrade. A manipulator opened orders with a volume exceeding the rest of the order book on the election day. Prices did not return to the previous level within the next few hours. Though the manipulator invested a lot of money, it was still less than one would pay for a prime time commercial.

Rhode and Strumpf (2008) and Luskin (2004) report a similar case in the 2004 U.S. Presidential Elections in favor of Kerry (against Bush). A manipulator spent $20.000 on the TradeSports prediction market to manipulate the price in the night of the first debate, knowing that the market has a huge impact on the public opinion. The contract for Bush was dropped from 54 to 10 percent. However, the price returned to 54 within 6 minutes. Two of this patterns happened directly after the second and

[5]This is also suggested by the results of Deck, Lin, and Porter (2013).

[6]See Section 2.2.5 for an explanation of market scoring rules.

[7]Endogenous incentives are provided by the markets, e.g. profits. Exogenous incentives are provided from outside the market, e.g. decisions (Hall, 2010).

third presidential debates, possibly "[...] to create the false impression of winning momentum for John Kerry" (Luskin, 2004, p. 1). Hardford (2007) and Wolfers (2007) suspect a case in the 2007/2008 U.S. Preliminaries in favor of Clinton. The last case was such a controversial one, that Koleman Strumpf, a well-known researcher on prediction markets, was not even sure, if this was really manipulation (Wolfers, 2007; Newman, 2010).

What many of these cases teach us and Rhode and Strumpf (2008) outline, is, that a manipulative attack may not only be successful if the prices were changed. If the intention was to generate media coverage or drive the momentum for a certain candidate, they may have been successful.

A very controversial discussion on manipulation was initiated by the proposed introduction of the Policy Analysis Market in 2003 that is described, among other, in Hanson, Oprea, and Porter (2006). The market was intended to forecast geopolitical events, such as the occurrence of terror strikes. Several authors indicated the hazards that may be caused by manipulation in this market.

TABLE 7.1: List of reported cases of manipulation/fraud in prediction markets.

Reference(s)	**Description & Findings**
Manipulation	
Rhode and Strumpf (2008) and Rhode and Strumpf (2004)	Analysis: Manipulation in the historical political markets in New York City between 1880 and 1944 in a comparison from betting odds and market prices. Though manipulative attempts can be supposed, they were not associated with large permanent effects in the prices.
Camerer (1998)	Experiment: Studied manipulation in horse-racing. Observed effects that very fast tended to a net effect close to zero.
Rhode and Strumpf (2008)	Experiment: Placed random bets on IEM (real money) in advance to the 2000 U.S. Presidential Elections that intended to mimic "insider" information. In total, it is suggested, that long-term market dynamics were not influenced.
Newman (2010)	Analysis: Prediction market IEM for the 2000 U.S. Presidential Elections. A group of trader raised the price for Buchanan.
Hanson, Oprea, and Porter (2006) and Wolfers and Zitzewitz (2004)	Discussion: A prediction market to forecast geo-political events, such as terror strikes, was supposed, but the project was canceled before its start due to several critics.
Rhode and Strumpf (2008), Luskin (2004), and Newman (2010)	Analysis: Manipulation by few large investors causes price drops in the Bush contract in TradeSports (real money) for the 2004 U.S. Presidential Elections. However, prices quickly (6 minutes) returned to their prior level and the trades were not profitable for the manipulators.

TABLE 7.1: List of reported cases of manipulation/fraud in prediction markets.

Reference(s)	**Description & Findings**
Wolfers and Leigh (2002)	Discussion: In the 2001 Australian Federal Election candidates bet on themselves at long odds in order to create buzz. No long-lived effects were observed (Wolfers and Zitzewitz, 2004, p. 119).(*Note: A check in the original source was conducted, but it was not successful to find the report in Wolfers and Leigh (2002))*
Hanson, Oprea, and Porter (2006)	Lab Experiment: Manipulators and normal traders forecast some figures, while all participants are aware of the incentives of the manipulators and the direction of action. In this setting, there were no effects on the market price.
Schröder (2009)	Analysis: Prediction market from Neue Zuercher Zeitung Online for the 2007 Switzerland National Elections. Media partners reported that the market was manipulated and that these participants were blocked. However, at least one trader still influenced the price of the Green Liberal party (consistently).
Hardford (2007), Wolfers (2007), and Newman (2010)	Analysis: Manipulation in favor of Clinton in the U.S. Preliminaries 2007/2008. This case was controversial, regarding if this was real manipulation.
Newman (2010)	Analysis: Prediction market Intrade for the 2008 U.S. Presidential Elections. Price of McCain contract was raised by 10 points by a "single large investor". Obama contract fall at the same time significantly. This also opened arbitrage profits when coupled to BetFair.
Cowgill (2006) from Rhode and Strumpf (2008)	Discussion: An employee admitted to attempting to manipulate Google's internal prediction market. With no effects.
Oprea et al. (2008)	Lab Experiment: Manipulated markets does not lead to worse forecasts by third parties.
Hanson and Oprea (2009)	Theoretical model: Manipulation has no long-lived effects on prices and even raise market efficiency due to its liquidity providing effect.
Veiga and Vorsatz (2010)	Lab Experiment: In few cases the presence of an uninformed robo trader helps a manipulator to successfully raise prices. In most cases there is no effect.
Deck, Lin, and Porter (2013)	Lab Experiment: Manipulated markets do lead to worse forecasts by third parties (same setting than Oprea et al. (2008), well-equipped manipulators with high incentives)
Jian and Sami (2012)	Lab Experiment: Several markets with two participants were tested. Manipulative strategies such as "bluffing" and "delaying" lead to slower price convergence, but no further differences.

TABLE 7.1: List of reported cases of manipulation/fraud in prediction markets.

Reference(s)	**Description & Findings**
Chen et al. (2015)	Lab Experiment: Markets with two participants were performed. If there are strong outside incentives it may lead to situations where manipulation is possible.
Buckley and Doyle (2017)	Field Experiment: 67 undergraduate students traded on a LMSR prediction markets (play money). The researchers used 16 manipulative accounts (4 per market) to drop the price. The effects were short-lived.
Huang and Shoham (2014)	Simulation: A prediction market with a market scoring rule is simulated. Agents know about the existence and direction of manipulation with various probability. If they are not aware of manipulators, the price may be influenced.
Rothschild and Sethi (2016)	Analysis: Manipulation by two large investors causes drops in the Obama contract in Intrade (real money) for the 2012 U.S. Presidential Elections in the night at the election day. Prices did not fully return to the previous level within the given time frame in which no new information were available. The cost for the attack was lower than that of a prime time commercial.
Fraud	
Bohm and Sonnegard (1999)	Analysis: Prediction market SEUPSM for the Swedish EU Referendum 1994. In a one-month competition shortly before the end fairly high monetary prizes were offered. Two coalitions formed in order to win the money (both won the prizes). "Shallow" markets were used for shifting money. No long-lived effects on accuracy.
Hansen, Schmidt, and Strobel (2004) and Newman (2010)	Analysis: Prediction markets Wahl$treet and Wahlboerse for 1999 Berlin State Election. The forecasts were published by local newspapers. A party animated their members to raise the forecast for their vote-share to influence public opinion. The attack was only successful in Wahl$treet, as in Wahlboerse there was a delay in account activation.
Blume, Luckner, and Weinhardt (2010) and Blume (2012)	Analysis: Prediction market STOCCER for the FIFA World Cup 2006. Several accounts were observed transferring money from a sacrificing account to a beneficiary. Assume temporary influences on market prices.
Schröder (2009) and Franke, Hoser, and Schröder (2008)	Analysis: Prediction Market PSM for the 2006 Baden-Wuerttemberg State Election. Two coalitions were observed.
George Mason University (2015)	Analysis: Prediction market SciCast to predict global developments 2012-2015. 14 fraudulent accounts were identified that intended to win the prizes. In addition bots are discussed to prevent cheating.

TABLE 7.1: List of reported cases of manipulation/fraud in prediction markets.

Reference(s)	Description & Findings
Kloker and Kranz (2017)	Analysis: Prediction market Kurspiloten to predict economic figures in 2011. Several fraudulent accounts were identified that intended to win the prizes.

Cases of Fraud

The probably best known case of fraud in prediction markets happened during the Berlin State Election 1999, reported by Hansen, Schmidt, and Strobel (2004). Two prediction markets were running (Wahl$treet and Wahlboerse) of which the results were published by local newspapers. The FDP, an economic liberal German political party, likely to fail the 5% hurdle in this specific election, animated their members to enter these markets in order to raise the forecast for their vote-share to influence public opinion. Though it may also be argued that this is just a further case of manipulation, all accounts/members were somehow controlled by the FDP headquarters and therefore can also be regarded as sibyls[8], which would be against the rules of the prediction markets. The attack was only successful in Wahl$treet, as in Wahlboerse there was a delay in account activation and, therefore, the attackers were not able to trade soon enough.

Real coalition building in order to win monetary prizes and with no interest in the accuracy or the result of the prediction is reported by Bohm and Sonnegard (1999). Bohm and Sonnegard (1999) operated the prediction market SEUPSM for the Swedish EU Referendum 1994. In a one-month competition, shortly before the end, fairly high monetary prizes were offered. Two coalitions formed in order to win the money of which both succeeded and won the first two prizes with a clear distance to the third place. Bohm and Sonnegard (1999) noted, that especially the shallow "BLANK" market was used for shifting money. A pattern for coalition trading is explained. The fraudulent attack had no long-lived effects on accuracy, though Bohm and Sonnegard (1999) emphasize that the rise of public interest in prediction markets makes them more prone to the influence of certain interest groups and that this may cause a real hazard to the accuracy of future prediction markets.

Blume, Luckner, and Weinhardt (2010) and Blume (2012) observed fraud in the prediction Market STOCCER for the FIFA World Cup 2006. Several (sibyl) accounts or coalitions were identified. All intended to win the prizes and had no interest in the market forecast. Irregular coalitions were also reported for the 2006 Baden-Wuerttemberg State Election by Schröder (2009) and Franke, Hoser, and Schröder (2008).

The George Mason University (2015) describes in their technology report in detail how 14 fraudulent accounts were detected in the SciCast prediction market. The SciCast prediction market is a combinatorial prediction market in which each question is represented by a LMSR market. The markets are connected in a Bayesian network and can therefore be answered conditioned by the outcomes of other markets. Several properties and features are described that may be used for a fraud detection algorithm, though such an algorithm was never implemented in SciCast. Instead, a visual collusion matrix was used as a starting point for manual fraud detection. The George Mason University (2015) emphasizes, as many other authors,

[8]Sibyls are defined in detail in Section 7.2.3.

that high monetary incentives are a major cause of fraud. In addition, it is suggested that the community of honest trader may be mobilized to observe and monitor rightful behavior. It is also noted that the publishing of blocked users immediately led to virtually no further fraud. Finally, the "intentionally losing" account is also characterized as a manipulative or fraudulent pattern.

Kloker and Kranz (2017) introduced the Fraud Cube to characterize manipulative and fraudulent attacks in order to help prediction market operators to get a better understanding of underlying motivations and common patterns. A rule based detection algorithm is suggested and demonstrated on historical data of the Kurspiloten prediction market. The Fraud Cube is explained in detail in Section 7.6.1.

Manipulation and Fraud in the FAZ.NET-Orakel

In the recent example of the FAZ.NET-Orakel, both was observed: Manipulation and fraud. Fraud was virtually existent in markets in all categories. After identifying and locking fraudulent and manipulative accounts, attempts were undertaken to contact the fraudulent traders in order to gain a better understanding of their thoughts. Some revealed interesting insights yet not described in literature. In one case, the operator was able to contact one fraudulent trader (multiple accounts) for a semi-structured phone interview (this case will be picked-up in Section 7.6.4). When asking for his motivation, the fraudulent trader stated that he perceived himself as a knowledgeable trader in the context of economics. However, due to the broad topical spectrum, he did not see his knowledge reflected in the ranking and therefore cheated in markets other than economic figures. Some other fraudulent accounts (though overall the minority) responded in an e-mail. One trader mentioned the fact, that he enjoyed seeing the figures and forecasts represent his expectations. Just for this reason he performed some trades to influence the displayed forecasts, even if they may have been obviously loosing.

Manipulation was observed for the 2017 German Federal Election only rarely until the last weeks, where suddenly a small group of participants started to raise the price (or prediction) of the AfD[9] in four large attempts. Three of the four attempts happened in the last week before the election, one large within the last eight hours (see Figure 7.3). The strategy was comparable in all four cases: A small group of traders started to buy everything on the sell side of the order book according to their purchasing power in the corresponding market and set high block-orders on the buy side at the same time. The last manipulative attempt was finally prevented by some truthful users that blocked the sell side with several large orders. As stated above, literature provides and explanation for this behavior: Manipulators want to present their preferred party as an "electable option" (Hansen, Schmidt, and Strobel, 2004; Rhode and Strumpf, 2008). They rely hereby on the idea that many people do not interpret the market predictions as what they are, but as results of election polls (Hansen, Schmidt, and Strobel, 2004). After a confrontation of these accounts by email, the manipulative accounts denied these accusations. However, a simple internet search of their user names quickly allowed to link at least two accounts directly to members of the aforementioned political party.

In theory manipulative attempts are short-lived. However, this does not justify a position that no countermeasures are necessary. Even temporary and short-lived changes in the prediction accuracy can lead to changes in beliefs of participants and (mis)reports of these false predictions in other media such as newspapers and social

[9] Alternative für Deutschland, German right-wing, conservative, anti-European political party.

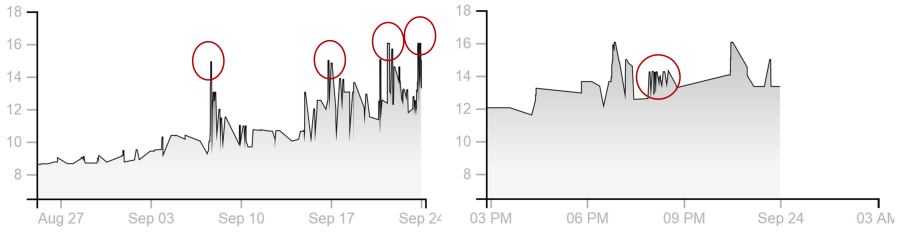


FIGURE 7.3: Manipulative attacks on the AfD stock on the FAZ.NET-Orakel. On the left side in a time span of the last four weeks, on the right side in a time span of the last eight hours.

media, which can influence voters and predictions[10]. In addition, some publications, demonstrated that given strong exogenous incentives, manipulation may even be finally successful (Deck, Lin, and Porter, 2013; Chen et al., 2015). In addition, fraudulent attempts, at least if they are perceived by the rightful traders, lead to a strong discouragement and frustration of other traders.

Current Detection Mechanisms

So far, only two implemented approaches of detection mechanisms are described in detail in literature, though there are reports on many others. Blume, Luckner, and Weinhardt (2010) identify two patterns that indicate fraud: The "ping-pong heuristic" and the "prominent edge heuristic". Blume, Luckner, and Weinhardt (2010) further developed an algorithm and figures to detect and quantify occurrences of those patterns. In a later publication, same researcher introduced a graph based implementation of these algorithms to enable prediction market operators to find fraud faster (Blume, 2012). In this implementation, Blume (2012) represented participants as nodes and transactions as edges. A more sophisticated approach was presented by Schröder (2009) and Franke, Hoser, and Schröder (2008). Schröder (2009) creates "Hermitian Matrices" and calculates "Eigenvectoren" and "Eigenspaces" in which anomalies and especially predefined patterns can be identified. Kloker and Kranz (2017) and the George Mason University (2015) advocate also to include account properties, such as IP-addresses or creation dates, into algorithms for manipulation and fraud detection.

Problem Formulation

To conclude the academic research, there is an urgent need of further detection and prevention mechanisms. Rule-based algorithms fail due to a lack of creativity and machine learning is not applicable due to sparse data. Especially regarding manipulation and controversial cases prediction market providers lack applicable instruments to deal with. There is also a lack of research regarding which design features hinder or promote fraud and manipulation (Rhode and Strumpf, 2008, p. 37), besides some papers that suggest the existence of strong outside incentives to promote manipulation.

The ADR paradigm starts by the problem formulation (Sein et al., 2011). To the best of author's knowledge the nature of the current problem, however, was yet not captured within an existing class of problems. The understanding of Dresch,

[10]This happened in the current case as well: http://www.faz.net/-gv6-922fq

Lacerda, and Antunes (2015) is applied with regard to what is considered "class of problems". The current problem has several properties: i) Manipulative and fraudulent traders always find new and creative strategies or generate a large understanding of the detection software to bypass detection. ii) There is not enough labeled data to train machine learning approaches. iii) There is an undesired behavior that is, however, not arguably or obviously against "the rules". iv) There is no "ground truth" available, as there may be cases of fraud or manipulation that were not detected yet or controversial regarding their (true) nature. Regarding some properties, the problem is actually very similar to those of sibyl detection in social networks or task approving in crowd-sourcing. Therefore, an appropriate class of problems would be *monitoring* and an appropriate subclass *monitoring in the context ambiguous and rare events*.

7.2.2 Kernel and Design Theories

Garcia-Molina et al. (2016) summarize different design elements that were frequently used in the field of *Data Crowdsourcing*. Data Crowdsourcing and the current problem (*monitoring in the context ambiguous and rare events*) are to some extent very similar, for which reason the collection of Garcia-Molina et al. (2016) can be considered a basic design theory for the current research project. Though several dimensions for classification are introduced and explained, some limitations regarding portability have to be considered. A clear distinction between the current problem and those addressed by Garcia-Molina et al. (2016) is the degree of creativity needed, the granularity of the tasks, and the incentives. Garcia-Molina et al. (2016) have a slight focus on labeling tasks, previously decomposed problems, and monetary incentives[11]. Within the class of data processing tasks, *filtering problems* are defined as one type of typical crowd-sourcing tasks. In general, according to Garcia-Molina et al. (2016) a data crowd-sourcing application has to fulfill four steps: (1) Prepare and initialize the data for the worker (define the tasks), (2) decompose and aggregate the data, (3) manage the worker (boredom, experience, etc.) and (4) optimize the task fulfillment or assignment with prior or external information. Design implications from the first step are to define an interface that supports the user in solving the problem and also in providing the solution. Design implications from the second step are to design an application that provides sufficient information or even trains the user for the task. Design implications from the fourth step are to include access to existent information from computed sources (e.g., the implemented fraud detection algorithm in the FAZ.NET-Orakel) or other users[12]. However, the current problem is different with regard to step two as it is not possible to identify a single transaction as fraudulent without its context of other transactions. Therefore, it is not possible to previously decompose the problem and offer the participants one transaction after each other, but to allow the participants to decompose and navigate through the problem by themselves at the granularity they need. In addition, in the understanding of Garcia-Molina et al. (2016, p. 903), all problems could "[...] in principle be solved by a single worker". In the current problem of fraud and manipulation, it was argued that the crowd is explicitly needed for the reason as it is assumed that a single worker does not have enough creativity to recognize all possible fraudulent

[11] Their focus is on data augmenting and data processing tasks in which the coordinator is a computer, workers are human, and the output is labeled data (training sets).

[12] Though, it has to be mentioned that crowd-sourcing settings are also subject to the bandwagon effect (Eickhoff, 2018).

and manipulative patterns. Therefore, fostering creativity is also a key challenge of the suggested tool.

Voigt, Niehaves, and Becker (2012) summarizes different design theories for Creativity Support Systems (CSS). The focus of these CSS is on information systems that support the creative process (e.g. idea creation). As the current problem also involves a certain degree of creativity, some of the facets or the kernel theories may add value for the design of the suggested artifact. Norman (1998) (The Design of Everyday Things) emphasizes that in order to foster creativity it is important to understand the user's task and its proceeding in solving it and to design the task/interface in such a way that the task is enjoyable. Design implications from this are the necessity of an understanding of the task, but even more, in order to foster creativity, to design the tool as enjoyable as possible, potentially "playful". Playfulness is also mentioned in the "Unified Design Theory for CSS" by Voigt, Niehaves, and Becker (2012)[13]. Further, CSS have to ensure comprehensibility (a rapid and clear understanding) and specialization (support the special purpose). Both confirm the design implications from step one, two and four of Garcia-Molina et al. (2016): The interface has to be created as simple as possible to ensure rapid understanding and task completion, add informative or training material and, add computed information where possible.

Candy (1998) identifies "[...] working with visual data such as images, drawings, sketches, diagrams, charts, graphs, [...]" as one important design feature for CSS. Further she mentions collaboration as important for creative work. Voigt, Niehaves, and Becker (2012, p. 170) derive from Candy and Edmonds (1995) three explicit design principles, one of relevance for us: "Exploration & Evaluation Support reflected in the activities of [...] examine data, which is supported by providing holistic views, multiple representations of data, visual data annotation, and concurrent processes [...]". The design implication of this DP in CSS is that the suggested tool has to enable its users to freely navigate and explore the data in order to promote creativity. At best, using a visualization that results in a "natural graphical interaction" (Voigt, Niehaves, and Becker, 2012, p. 170). The self-selected granularity, though this effect is not expected to be to large, in addition, may result in a balancing of skill and challenge level with respect to the *Flow Theory* (Nakamura and Csikszentmihalyi, 2014) and therefore foster performance and motivation. This effect is for example observed in online games (Hsu and Lu, 2004).

Regarding the continuous operation of such a tool one has to consider the incentives. Garcia-Molina et al. (2016) focus on monetary incentives, but also define the other options on the incentives dimension: Entertainment, learning, philanthropy, and "hidden". There exist plenty of kernel theories regarding monetary incentives and task performance or monetary incentives and creativity. And, associated, a lively discussion. Erat and Gneezy (2016) find that monetary incentives are more likely to reduce creativity. The *Motivation Crowding Theory* (Frey and Jegen, 2001) also suggests that monetary (extrinsic) incentives are "crowding-out" intrinsic motivation. Though learning or philanthropy cannot be expected and the operator does not have the financial power to subsidize the tool in the long run, it has to be fully relied on entertainment[14]. Further sources of motivation may be social components (Online Community Building, Kim (2000)) or, in current special case, the "hunger

[13] "Playfulness is the property of a tool to encourage unfettered trialability in design, helping the user to push intermediate solutions to final results iteratively." (Voigt, Niehaves, and Becker, 2012, p. 159)

[14] "Hidden" is not defined further in Garcia-Molina et al. (2016) and the it is not finally clear, what exactly is meant by it.

for justice". Experience showed that many participants wrote e-mails to the administrator, when they perceived that other participants cheated. In Section 5, it was summarized how social elements can increase motivation (and retention). Although the remarks there were made in the context of Real-time Delphi, many of the underlying kernel theories may be true here as well. In addition, it is also perceived to be relevant to ensure anonymity among the participants in this context to suppress hostilities. The design implication thereof is to implement spaces for discussions and add labeling (e.g. likes and dislikes) in the suggested tool.

7.2.3 Crowd-sourced Fraud Detection in other Contexts

Most research on fraud detection in the context of crowd sourcing and utilizing the crowd was done regarding how to validate submitted work from crowd workers, such as in Amazon Mechanical Turk or comparable platforms (Xintong et al., 2014). As one of the earliest approaches, Hirth, Hoßfeld, and Tran-Gia (2013) suggested two models to perform this task by the crowd, of which only the second approach is somehow related to current problem statement. A task, after submitted is split into parts and each part is forwarded to other crowd workers to rate the submitted work (to given criteria). The main task is accepted, if the majority of rates have been positive. This approach, however, does not require any creativity in finding fraudulent patterns, as they only rate according to given criteria and there is no need to consider sequential actions. Many other approaches to detect invalid work or invalid tasks followed (Baba et al., 2013; Kittur et al., 2013).

Almendra and Schwabe (2009) was one of the first to implement a pilot study where the crowd (non-specialized people) was used to detect fraudulent accounts in an online auction context. They were shown different seller profiles as they may be on an online auction platform and had to decide which profiles may be fraudsters. Here the task for the crowd workers included much more creativity and required them to consider different attributes simultaneously. While human computation delivered very good results in simple tasks, it is stated that fraud detection "[...] is certainly more difficult than giving descriptive labels for images" (Almendra and Schwabe, 2009, p. 302). In regard to their first question, whether this task is feasible at all for the crowd, the results were very promising, though not yet generalizable. The second question put by Almendra and Schwabe (2009), whether the cost of this method (reward for the crowd) is below the potential savings, is, however, subject to further research.

In the context of "citizen sensing" crowd-sourced fraud detection was used, e.g., during elections. In Bader (2013), the crowd was used to collect information about fraud during the 2011-2012 Russian Election. The Russian people were asked to report all kind of fraudulent actions they observed during the election and to collect them on an online portal called "Karta Narusheniy". There, the reported cases were categorized and visualized. Shayo and Kersting (2017) applied a similar method in the 2015 Tanzanian Election.

Another stream of research in this context deals with the identification of web threads. In an experiment, Moore and Clayton (2008) evaluated the wisdom of crowds regarding the detection of phishing websites. Sharifi, Fink, and Carbonell (2011) implemented the idea in a browser plugin. Besides phishing websites the plugin also means to detect all kind of web threats. Besides the votes and comments of the crowd, this data is combined with machine learning approaches and natural-language processing. The plugin featured a bookmarking system with question-answering functionality to promote wider user participation.

Frequently the idea to use the crowd to detect misconducting behavior is discussed in the context of online social networks. Ghosh, Kale, and McAfee (2011), e.g., are discussing this issue and corresponding problems. The work also mentions the problem when not all raters of the crowd are trustworthy and suggests an algorithm to solve this problem. This problem is indeed universal to many crowd-sourced detection tasks. While Ghosh, Kale, and McAfee (2011) is focusing on abusive content in online social networks, Wang et al. (2013) apply a method to use crowd-sourcing to detect sibyl accounts. Sibyls are fake accounts in online (social) networks and regularly used to spread spam, malware, (false) information, or to push topics. In recent years, sibyl accounts became more and more professional. In this context, the study of Wang et al. (2013) compared the detection quality of the crowd (Amazon Mechanical Turk) compared to experts. Though the result of the crowd was varying but good, experts produced near-optimal results. Wang et al. (2013), therefore, showed that the crowd is capable to evaluate fraudulent behavior that includes a certain degree of creativity on the fraudsters' side and without a "clear schema" to check the object of interest. However, a combined approach with the crowd and experts would yield superior performance, which, subsequently, is introduced by Wang et al. (2013).

There is some literature, how the crowd is utilized to create fake reviews for products. Only few papers deal with the use of the crowd to detect such reviews[15]. Harris (2012) compared a crowd-based approach versus machine-based approaches and found that "[...] the combination of human-based assessment methods with easily-obtained statistical information generated from the review text outperforms detection methods using human assessors alone" (Harris, 2012, p. 87).

If the focus is only put on contexts, where financial transactions have to be monitored, there have also been several applications of crowd-sourced fraud detection.

Theodoulidis and Diaz (2012) advocates "crowd monitoring" to ensure compliance in high frequency markets. Though the method itself is not described in detail (and not validated), the work mentions some interesting assumed advantages, among others: "Consequently such crowd-monitoring agents will be capable to analyze more complex scenarios, including possible new HFT [(high frequency trading)] based manipulations, and or cross-border and cross-asset ones" (Theodoulidis and Diaz, 2012, p. 11).

Matti, Zhu, and Xu (2014) developed a tool that identifies potential fraudulent persons via twitter by checking all new tweets matching a list of words (e.g., bankruptcy, fraud, etc.) and rechecks these persons in transaction databases of, e.g., credit card institutions. The Guardian, a British online magazine, created a public database of more than 700,000 expense claims of the members of the UK parliament to search for the public (Kramer Mayer and Hinton, 2010). Some 20,000 participants joined the search and fueled a national scandal.

Noteworthy is also the patent by Anderson and Ross (2014). A tool for "Collaborative Fraud Determination and Prevention" in the context of financial transaction is suggested. Several "reviewing entities" (e.g., merchants) can pool and share customer and transaction data to a central database/platform. They share their data already labelled in fraudulent and non-fraudulent transactions that is joined according to the customer with the provided data of other reviewing entities. Future transactions can be validated based on all data using a "fraud determination query" that takes the information of all merchants into account. An interesting feature is that the reviewing entities can be rated by other reviewing entities along several dimensions,

[15]Though the companies such as Amazon are very likely to employ humans for such tasks.

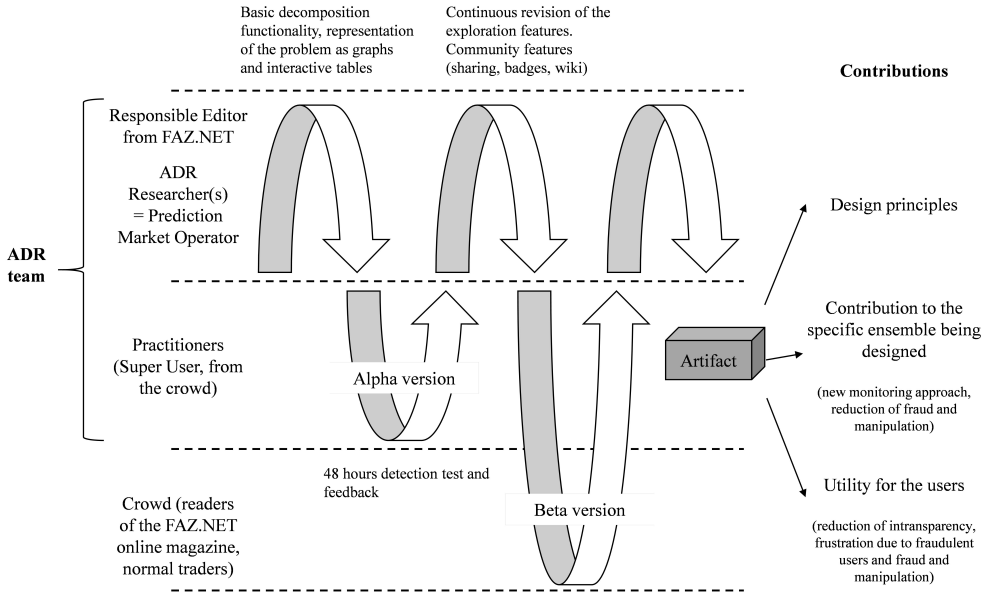


FIGURE 7.4: Schema for IT-Dominant BIE for current ADR research project. Adapted from Sein et al. (2011, p. 42).

among others reliability and trustworthiness. So, in this case the data and the task of quality control over the participants is crowd-sourced – not the "fraud determination" itself. However, the platform also contains some functionality to generate some reports automatically.

7.3 Method

> Contents of this section are in part adopted from Kloker et al. (N.D.).
> See Section A.1 for further details.

Based on the considerations in the preceding Section 7.2, an artifact for crowd-based detection of manipulation and fraud was developed following the ADR methodology.

In this section the course of events for the BIE (Building, Intervention, and Evaluation) phase is presented. In addition, the context (partners and organizations) of current research project and relevant methodical considerations will be defined and explained. The implementation details of the artifact will be addressed in Section 7.4.

The IT-dominant schema was applied for the current project, as the end-users of the tool are not within the researchers, nor the partner's organization (IISM, FAZ-.NET), but customers at the same time. Following the suggestion of the IT-dominant schema, two cycles were performed. The adapted BIE process is illustrated in Figure 7.4.

The FAZ.NET-Orakel is described thoroughly in Section 3.2, for which reason this section will only briefly repeat the most important points. The prediction market FAZ.NET-Orakel features predictions regarding politics, economic figures, sports, and other current events of interest. The FAZ.NET-Orakel is online since March

2017 and more than 1600 users traded at least one time. The FAZ.NET-Orakel receives high public visibility, as the FAZ.NET regularly publishes articles on the current forecasting competitions and results. This is especially true for articles on elections. If one searches the internet regarding the 2017 German Federal Election and the FAZ.NET-Orakel one finds both, announcements from people to manipulate the prices of the AfD stock[16] or calls to prevent predictions to come true[17]. The prediction market is publicly available to all readers of the FAZ.NET and beyond, though trading is tied to a FAZ.NET account that can be created free by any holder of an e-mail address. The registration and verification process is performed by a subcontractor of FAZ.NET, for which reason the market operator does not have any influence on the technical implementation and process. This subcontractor, however, does not filter spam- or trash-mail addresses or verify anything beyond the e-mail address[18]. As the prediction market is operated within the brand FAZ.NET, the design is adapted to the corporate design. In addition, a high importance is given to an intuitive usage and simple market design. The ranking features an algorithm that allows participants to enter the competition to a later point of time and still have reasonable chances to end up in a high rank (if he performs very good). As described in Section 5 the market also features Delphi-studies.

The FAZ.NET-Orakel features a state-of-the-art fraud detection algorithm derived from Blume, Luckner, and Weinhardt (2010) that monitors every transaction for the "Ping-pong" pattern and therefore identifies obviously loosing transactions. The algorithm distributes "suspicious points" to accounts for suspicious transactions (as suggested by Kloker and Kranz (2017)). If an account receives more than three points within 24 hours, it is displayed a warning message, saying that the account shows irregular trading behavior and encouraging the holder to reconsider his transactions (see Section 7.6.4 for more details). If an account receives more than five suspicious points, it is immediately locked. In addition, the market operator checks once a week if different accounts used same IP-addresses.

The ADR team consists in the inner circle of the ADR researcher, developer and operator of the FAZ.NET-Orakel (affiliated with a research institute, IISM) and a responsible editor/journalist (affiliated with the Frankfurter Allgemeine Zeitung, FAZ.NET). The researcher is regularly in exchange with some super-users of the FAZ.NET-Orakel, who can be regarded as "the practitioners", as they are both highly knowledgeable in the matter and directly report to the researcher. All other participants are referred to as the end-users or the crowd (though the super-users are actually also part of the crowd).

These super-users are the "initial knowledge-creation targets". In several e-mails and phone calls, the super-users complained about unusual activities on the prediction market. Some of these cases were detected by the existing algorithm, but many not. For this reason, the practitioners offered to support the search for manipulative and fraudulent patterns, if they get access to the underlying data. This communication at the interface between the researchers and practitioners was the initiation of the project.

The following section describes the design process of *guided emergence* of the suggested IT artifact (Sein et al., 2011).

[16] https://www.youtube.com/watch?v=9xxfetB0Yac, start from 2:25 min

[17] https://www.myheimat.de/uelzen/politik/faznet-orakel-afd-bei-mehr-als-13-prozent-d2833112.html

[18] This is important, as many suggested to the operator, to filter for such trash-mails or to verify natural persons. Indeed, some fraudulent cases were performed by such accounts. However, several these often used very basic strategies.

7.4 Building, Intervention and Evaluation

> Contents of this section are in part adopted from Kloker et al. (N.D.).
> See Section A.1 for further details.

7.4.1 Alpha Cycle: Implementation of the detection component

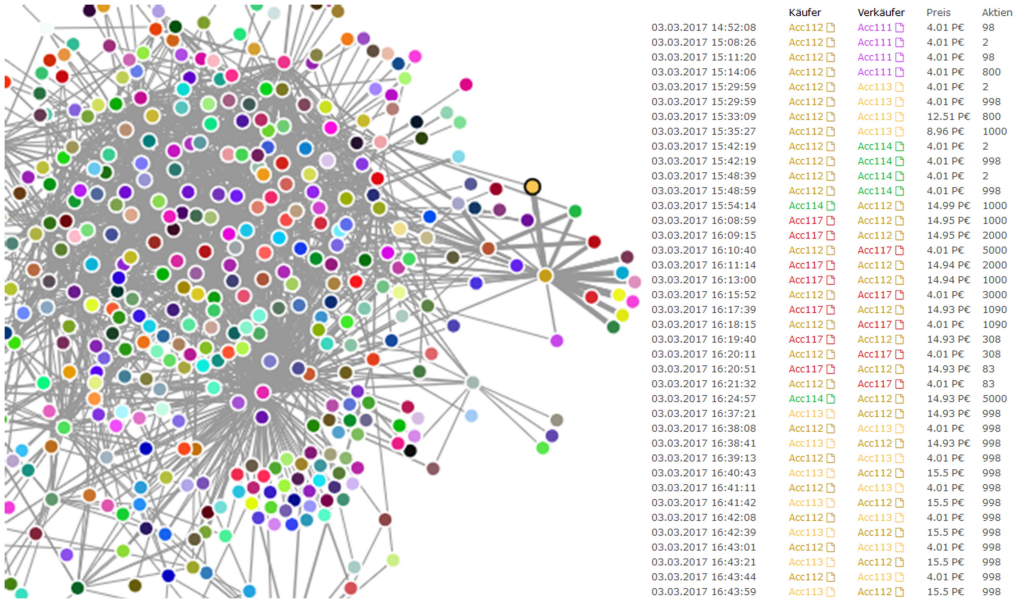
Initiated by the communication between super-users and the researcher regarding still prevalent cases of fraud and their suggestion to support search, a first version of the artifact for crowd-based manipulation and fraud detection was developed. This artifact featured basic functionality for decomposition, exploration, detection, and reporting of fraudulent cases. The artifact was presented to selected super-users in the context of a small pretest (limited scope, only one market). As the prediction market is accessible by all readers of the FAZ.NET online magazine, it is of utmost importance to publish artifacts for a broad audience only in a finished state.

Building

In order to enable the crowd to find fraudulent acts, it is necessary to present the current transactions of the market in such a way that they are able to understand and break down (decompose) the problem of the search for manipulation and fraud. As a reference point to start designing the artifact, the representation of the trading behavior from Blume (2012) was applied. In Blume (2012) a graph based visualization of the trading activity is introduced. Each trader is represented by a node. Transactions between two traders are represented by directed edges between the nodes. Edges between the same nodes and with identical directions are aggregated. The strength of the edge indicates the volume of the transactions (a strong edge indicates high volume transfer, a thin edge indicates low volume transfer). The suggested artifact also utilizes this representation. It goes, however, without the directed edges and aggregates all transactions, independent of direction, between two nodes. To keep the graph simple, it only shows the labels when the mouse is moved over a node. For an easier reference of the nodes in the graph, each node is displayed in a unique color. Nodes can be moved by dragging while the other nodes and edges move simultaneously to ensure that all nodes remain visible and are arranged in a reasonable way. In detail, the arrangement of the graph is determined by three competing forces that strive for equilibrium: (1) All nodes are drawn to the center of the graph, (2) all nodes repel each other, and (3) each edge tries to minimize its length. The nodes move (not only hide or appear) and have a kind of elastic behavior, so that moving the nodes results in a kind of playful behavior (to foster creativity according to the kernel theories on creativity).

In addition to the graph, each transaction is displayed in a table below that contains the following information: (1) The market in which the transaction took place, (2) the date and time of the transaction, (3) the buyer's name (uniquely colored) and the corresponding order, (4) the seller's name (uniquely colored) and the corresponding order, (5) the execution price of the transaction, (6) the volume transferred, and (7) a flag and a number indicating how often the transaction has already been identified as potentially fraudulent or manipulative. The account names are displayed in the same unique color that is used in the graph, so an easy cross-referencing is allowed. The table does also show bundle trades of single accounts.

The account or node names are assembled by the prefix "Acc" and an ascending number, e.g. "Acc293". The ascending number indicates the order of the accounts



	Käufer	Verkäufer	Preis	Aktien
03.03.2017 14:52:08	Acc112	Acc111	4.01 P€	98
03.03.2017 15:08:26	Acc112	Acc111	4.01 P€	2
03.03.2017 15:11:20	Acc112	Acc111	4.01 P€	98
03.03.2017 15:14:06	Acc112	Acc111	4.01 P€	800
03.03.2017 15:29:59	Acc112	Acc113	4.01 P€	2
03.03.2017 15:29:59	Acc112	Acc113	4.01 P€	998
03.03.2017 15:33:09	Acc112	Acc113	12.51 P€	800
03.03.2017 15:35:27	Acc112	Acc113	8.96 P€	1000
03.03.2017 15:42:19	Acc112	Acc114	4.01 P€	2
03.03.2017 15:42:19	Acc112	Acc114	4.01 P€	998
03.03.2017 15:48:39	Acc112	Acc114	4.01 P€	2
03.03.2017 15:48:59	Acc112	Acc114	4.01 P€	998
03.03.2017 15:54:14	Acc114	Acc112	14.99 P€	1000
03.03.2017 16:08:59	Acc117	Acc112	14.95 P€	1000
03.03.2017 16:09:15	Acc117	Acc112	14.95 P€	2000
03.03.2017 16:10:40	Acc112	Acc117	4.01 P€	5000
03.03.2017 16:11:14	Acc117	Acc112	14.94 P€	2000
03.03.2017 16:13:00	Acc117	Acc112	14.94 P€	1000
03.03.2017 16:15:52	Acc112	Acc117	4.01 P€	3000
03.03.2017 16:17:39	Acc117	Acc112	14.93 P€	1090
03.03.2017 16:18:15	Acc112	Acc117	4.01 P€	1090
03.03.2017 16:19:40	Acc117	Acc112	14.93 P€	308
03.03.2017 16:20:11	Acc112	Acc117	4.01 P€	308
03.03.2017 16:20:51	Acc117	Acc112	14.93 P€	83
03.03.2017 16:21:32	Acc112	Acc117	4.01 P€	83
03.03.2017 16:24:57	Acc114	Acc112	14.93 P€	5000
03.03.2017 16:37:21	Acc113	Acc112	14.93 P€	998
03.03.2017 16:38:08	Acc112	Acc113	4.01 P€	998
03.03.2017 16:38:41	Acc113	Acc112	14.93 P€	998
03.03.2017 16:39:13	Acc112	Acc113	4.01 P€	998
03.03.2017 16:40:43	Acc113	Acc112	15.5 P€	998
03.03.2017 16:41:11	Acc112	Acc113	4.01 P€	998
03.03.2017 16:41:42	Acc113	Acc112	15.5 P€	998
03.03.2017 16:42:08	Acc112	Acc113	4.01 P€	998
03.03.2017 16:42:39	Acc113	Acc112	15.5 P€	998
03.03.2017 16:43:01	Acc112	Acc113	4.01 P€	998
03.03.2017 16:43:21	Acc113	Acc112	15.5 P€	998
03.03.2017 16:43:44	Acc112	Acc113	4.01 P€	998
03.03.2017 16:43:59	Acc113	Acc112	15.5 P€	998

FIGURE 7.5: Tool for crowd-based manipulation and fraud detection. The transaction overview on the left-hand side. An excerpt from the transaction history on the right-hand side. Each node identifies a trader, each edge indicates transactions between two traders. The strength of these edges indicates the quantity. Each trader has its own color. The map shows a suspicious case (at the right-hand side of the large point cloud) where some nodes traded very high volume exclusively with only one other node.

in which they entered the market (first transaction). This naming ensures account privacy (as the first 48 hours of trading are never displayed as long as the market is running). Suspicious transactions (not suspicious accounts) are reported by a small button next to the transaction in the table. Reporting a transaction always includes to provide a short comment on what is perceived as "suspicious".

The tool allows the subset of the complete graph based on any combination of markets, accounts and dates[19]. At the selection of accounts it can be distinguished between a mode where all transactions with selected accounts are displayed and a mode where only transactions between selected accounts are displayed. For a closed market the tool shows the complete transaction history. For a running market the tool shows only the last 24 hours[20] of trading and nothing within the first 48 hours after the opening. The reason therefore is to prevent the possibility to track single real accounts by combining information of the ranking and the transactions. This, however, induces the risk that manipulative and fraudulent transactions within the first at least 24 hours are not discovered before the closing of the market. Accounts that were previously identified as manipulators or fraudsters are indicated by a small icon next to the pseudonym of the account.

Exemplary user interfaces of the tool are illustrated in Figure 7.5 and Figure 7.6.

[19]The selection of the specific dates was implemented after the first 48 hours of publishing as the result of feedback of the super-users.

[20]Within the first 48 hours of the evaluation of the alpha cycle. Later, this figure was set to one week and only adapted after feedback of the super-users.

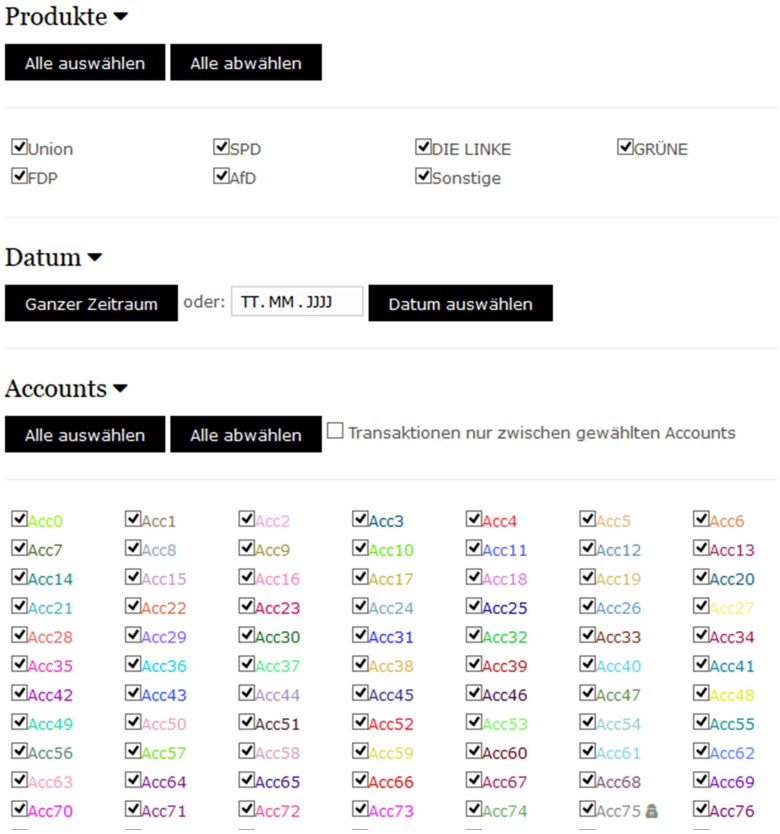


FIGURE 7.6: Tool for crowd-based manipulation and fraud detection. The tool allowed to subset the transactions based on any combination of markets, accounts and dates. It was also possible to decide if the graph should display all edges regarding selected accounts or only between the selected accounts.

A link to the tool was showed below every market, leading directly to the configured tool for this market. The tool is only accessible to registered and logged-in users.

Intervention

The implemented alpha version artifact was published on the 27th September 2017, 11:00 am on the prediction market.

To validate the general usability and expediency a study was performed within the first 48 hours after the go-live of the tool. The super-users were informed by a short note in the thread of the internal forum, where current cases of manipulation were discussed previously and on the news channel of the FAZ.NET-Orakel. No concrete work order was provided. No incentives were offered. So, the interaction of the traders with the tool was solely motivated by themselves.

The tool provided the full functionality, besides the indication of already detected and locked fraudulent or manipulative accounts and the selection of the time span. The traders had the chance to communicate with each other in the internal forum in the thread where the tool was announced. In addition, they had the opportunity to communicate to the administrator by e-mail. During the 48 hours the traders were provided with feedback on a regular basis, if their suspected and reported certain accounts were a hit or not. After the 48 hours the study was closed and all manipulative and fraudulent accounts detected by the crowd, as well as those that remained undetected but were known to the operator, became indicated.

Several kernel theories emphasized the importance of an intuitive and easy usage of the tool, for which reason no support was provided to the user at this time (e.g., in form of a manual), in order to see if they understand, how to use it.

Evaluation

The evaluation of an instrument for manipulation and fraud detection and monitoring is especially difficult, as it is not possible to intentionally induce and observe the phenomenon realistically in live data (with the creativity of real manipulators) and there is no "ground truth" of all manipulative actions in real-world (historical) data. At the time of publishing more than approx. 55,000 historical transactions were carried out on the FAZ.NET-Orakel, within the previous six months, and could be accessed and explored with the tool.

It was decided to evaluate the expediency of the tool on the historical transactions. That is why the tool was introduced in the first 48 hours without indicating the accounts that are already known for fraud or manipulation. The crowd was unaware of this move and they were not provided any incentive to look for fraud. All accesses on the tool, reported cases, and comments regarding the tool were tracked. As a benchmark to assess the detection quality of the crowd in the first 48 hours, the list of known cases of fraud and manipulation by the operator was utilized. This list consisted of all accounts detected by the operator using the detection algorithm and also other tools (see Section 7.3).

The results of the first 48 hours are reported in Table 7.2. The results show that the crowd was enabled to explore the large data set within a very short time and find more than two third of all previously identified cases very fast. Even more encouraging is the fact that two yet unknown cases were identified.

TABLE 7.2: Summary first 48 hours

Previously identified fraudulent accounts	44
Accounts reported (of known)	30
Accounts reported (of unknown)	2
Comments	14
Flags	8
Active users	15
Accesses of the tool	146

After the introduction of the indicators for banned accounts, another yet unknown fraudulent account, four suspicious accounts and ten accounts that showed at least strange behavior were reported in the historical data[21].

The results also show the users did not really use the flags, but the forum to report the cases. The participants were encouraged to give feedback via the forum or give suggestions if they miss features. The feedback was continuously positive.

Reflection and Learning

To begin with, it can be summarized that the tool itself fulfilled its basic need, to enable the crowd to explore and decompose the transactions in order to find fraudulent patterns. Also new and unknown cases were identified. Thereof, it can be derived that the representation in form of graphs and tables was applicable. The fact that almost no flags were used, but all cases were directly reported in the forum, showed that users have a need to communicate and exchange on the cases. Though it cannot be concluded at this point if this is due to the need for information exchange or social exchange (to receive social reputation), or both. It also seemed that the users had problems in referencing and describing single cases in order to enable others to retrace and find them. The flags seemed not to fulfill these needs. Feedback of the super-users stated that the time span at the opening of a market in which data was hidden is too long and that a selection criteria for a time span was missing.

7.4.2 Beta Cycle: Reshaping of the reporting and adding of motivational features

Within the first six weeks, usage figures were falling. This was not surprising, as it was to assume that participants were finished with the historical data and thereafter they only focused on new transactions, which, however, needed much less time and accesses. For the operator it is only important that the tool is continuously used and so the prediction market is continuously monitored. For this reason the tool was reshaped in a beta cycle according to the learning from the alpha cycle. In addition, further features were added to ensure (and further understand) long-term motivation derived from kernel theories in Section 7.2.

Building

First, the reporting component was revised. If a case is reported using the flag button a thread is automatically created in a dedicated forum for cases of manipulation and fraud. This thread includes a direct reference that jumps on a click to the case in

[21] The missing 14 of 44 previously identified accounts were not reported anymore, as they were then indicated and there was therefore no more reason to report these cases

the tool and pre-selects the suspicious transaction. Like this it is intended to facilitate the retracing of a single case that the users previously tried to explain in words in the alpha cycle. The forum featured basic features, such as likes and dislikes. The user names of the reporting accounts was not shown to the public to prevent revenge actions. However, each name of the commentators were shown. The dedicated forum that was later extended by a voting mechanism (see the gamma cycle) is illustrated in Figure 7.8.

Second, the "Sheriff"-badge was introduced. Participants that reported a suspicious case received a "Sheriff"-badge for two weeks. The badge was displayed in the form of a little star directly behind the users name throughout the prediction market and forums. A tool-tip text explained the meaning of the star/badge. The feature enables users to build up reputation within the system. This was already suggested to raise long-term motivation in other contexts (e.g., see Section 5).

Third, a wiki was introduced in order to capture known strategies and to maintain a manual for new users how fraud can be identified. The wiki started with one exemplary entry on the "ping-pong" strategy. Participants were encouraged to provide articles on further patterns.

Intervention

Before the second intervention, several further small improvements were implemented, especially regarding the speed and navigation through the data.

The update of the tool with the new features of the beta cycle (described in the subsection above), then took place on the 15th December 2017. The new features were not announced specifically.

Evaluation

To derive statements on the success of the tool and intervention, the time from the introduction of the tool (27th September 2017) until 12th February 2018 is evaluated. Figure 7.7 shows the overall usage and already suggests that the tool usage is mainly driven by a single super-user ("SuperUser1"). At all, 39 users used the tool. This figure includes 18 power users (continuously active trader on the prediction market[22]) and 21 casual users of the prediction market. The tool was used 1762 times (opposed to 37.012 trades of the corresponding users in the corresponding markets)[23]. As mentioned above, the decrease of usage in the first weeks can be explained by the fact that the historical transactions were finished. In addition, it has to be considered that the overall usage of the prediction market recorded a decrease in interest after the 2017 German Federal Election.

In this light, it is evaluated what motivates users to use the tool.

First, a significant, moderate, positive correlation (r=.526***)[24] between the individual interest in a market (measured by trades) and the engagement in the tool (measured by accesses to the tool) is found. This basically means that users that are interested in a specific market/topic are also more likely to engage in manipulation monitoring and detection in this market.

A very strong interest at the beginning is observed, which may be also attributed to the novelty effect. Second, it is found that many periods of "high usage" can be

[22] More than 1000 orders.

[23] All this figures include the administrator's accesses, as he also used the tool for fraud detection. The administrator is referred to as a power user, though he did not trade. 154 accesses were issued by the administrator.

[24] Significance codes: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$

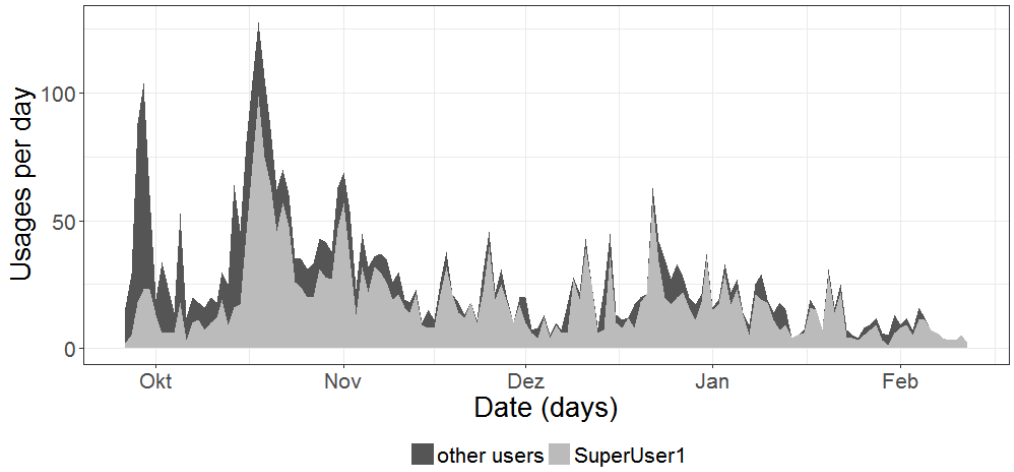


FIGURE 7.7: Comparison usages by the user "SuperUser1" and all other users.

directly related to the approaching closing of popular markets[25]. This suggests that the users monitor for manipulation, when they want to ensure that the markets will close correctly and the prices are distributed in a just manner.

The period of low usage at the beginning of February may be explained by the fact that almost no markets of public interest closed during this period.

Reflection and Learning

The first learning from the beta cycle is that the number of active users maintained at a low level. Many participants seemed not to be interested in engaging in monitoring efforts (that are not linked to any incentives). Nevertheless, the monitoring was done reliably by the few users. The active users reported six further suspicious cases (all of them not reported by the algorithm) after the introduction of the new reporting features (beta cycle) and, thereby, used the new functionality. Therefore, it is to assume that the reporting and referencing component of the artifact now fulfilled their needs, though the figure may yet not be statistically significant. However, all six reported cases remained suspicious due to a lack of evidence. As far as this can be assessed by the operator, there was no further fraudulent or manipulative approach on the platform within the evaluated time frame. It is not known if this is caused by a discouraging effect of the tool, low interest, or for the reason that at the moment no competitions with "high stakes" were active. It also has to be considered that the usages of the tool may be driven by the current overall interest in the prediction market or single competitions and the overall impression by the users that this market is manipulated or cheated. If this impression is not given for the moment, no high activity in the tool has to be expected. In addition, no feedback on the badges was received and no further articles were proposed for the wiki. At least for the latter, therefore, it can be assumed that users did not rely on this information or did not want to share this knowledge.

Only one user was active throughout the evaluated time period and functioned as a kind of continuous monitor. In a personal interview, however, he stated that

[25] Mid of October: Niedersachsen State election, mid of December: Bundesliga winter championship, end of December: Vierschanzentournee, end of January: Grammy Awards

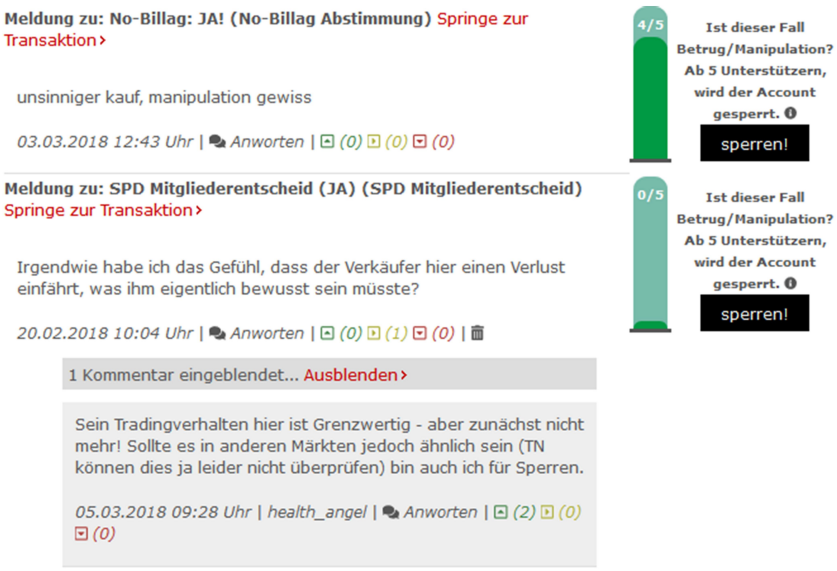


FIGURE 7.8: Voting tool to decide on ambiguous cases. The user interface shows two reported cases, of which the second is commented. The red link next to the subject of each reported case triggers a jump to the transaction in the transaction table in the tool. On the right-hand side of each reported case, users can vote to lock the traders involved in the reported transactions. Four users voted that the first reported case is fraud or manipulation, while the second is not perceived to be fraudulent.

he had used the tool besides of manipulation detection for other purposes. These were, e.g., to get an understanding of the strategies of several suspicious accounts that he perceived to be harmful to the market (e.g., overbid respectively underbid both sides of the order book in order to gain windfall profits). The user used this knowledge in addition to develop preventive trading strategies.

As the monitoring efforts seemed to increase right before the closing dates of the market, the number of visible days was set to the maximum number (current active time minus the first 24 hours). This was done to ensure that the habit of only looking right before closing does not harm detection in the early phase of a market.

7.4.3 Suggestion for Gamma Cycle: Voting Mechanism on Manipulative and Fraudulent Cases

A third ADR cycle was just initiated to address the second research question put in Section 7.1 ((II) How to decide on ambiguous manipulative and fraudulent cases that are not obviously against the rules.), although this cycle was not performed in the scope of the original ADR project anymore. To address the problem of ambiguous cases a further feature is suggested that allows to vote on reported cases. A voting tool was implemented that accompanied each thread in the dedicated forum, which was previously introduced in the beta cycle. The implementation is illustrated in Figure 7.8.

The voting tool was introduced on the 20th February 2018 and announced in the news feed of the FAZ.NET-Orakel. Since then the tool was used to decide on six new reported cases. It was defined that five users have to assess a reported case as

manipulative or fraudulent in order to block the related accounts. None of these cases were finally assessed as intentionally problematic. The operator also assessed none of these cases as manipulation or fraud. Though the usage figures are still low due to the small time horizon, it can already be concluded that the tool has been used to vote and also encouraged a discussion on the reported cases.

A detailed and quantitative evaluation is part of future work.

7.5 Discussion and Conclusion on the Crowd-based Detection

Stage 4 of ADR defines the steps from the results to formalization and learning with regard to the class of problems. The class of problems *"monitoring in the context of ambiguous and rare events"*, defined in Section 7.2, was addressed. With the crowd-based detection tool, it was demonstrated that a crowd is capable to perform such a monitoring task in general in the context of the presented prediction market. In the introduced case of the FAZ.NET-Orakel, the crowd was able to decompose a relatively large problem and detect even complex and unknown fraudulent and manipulative strategies.

Several design implications were derived from related literature, their applicability demonstrated in the current project, and evaluated in the implemented artifact. After evaluation and reflection of each design element and the acquired experience, it is now possible to formulate following DPs as potentially beneficial for other projects that address the same class of problem:

Design Principle 1: *Help the crowd in order to freely decompose the problem.*
A key feature of such a tool is the decomposition of the problem and the free navigation through the data. This may include natural navigation using visual graphs, but also universal color codes to identify single accounts/entities etc. throughout the application.

Design Principle 2: *Enable exploration to enable your crowd to think besides known patterns.*
To find creative patterns and strategies, it is important not to enforce basic schemes of data subsets, but to allow users to select and reselect subsets according to their own ideas.

Design Principle 3: *Enable references on and exchange of detected cases.*
Detected cases have to be made accessible to all participants. Software should support the reference and resubmission of suspicious cases. Other users need to be able to retrace and discuss each case within the community.

Design Principle 4: *Create value for your crowd by the artifact.*
It is important to understand what encourages your crowd. In current case it was the frustration regarding the in-transparency of the prediction market and its transactions and the perceived manipulation. Interviews with the users shed light that though the artifact was playful and well-designed, the users only used it to create justice and find other users guilty, if they perceived harmful behavior.

Following these DPs it was possible to implement an IT artifact that enabled the crowd to monitor the prediction market successfully for fraud and manipulation. Even complex patterns were found and reported, which is clearly dominant to existing algorithmic approaches in the context of prediction markets.

However, several limitations have to be mentioned. First, prediction markets are a traditional crowd-based approach for which reason it was to expect that there is a crowd which is potentially willing to engage. It is yet not finally clear, how the

design elements of the prediction market influence the usage of the tool and vice versa. Second, the current case showed a high dependency of the success of the tool on one single user. It is not surprising that out of about 20 active users only one contributes the most of the content, as this is the case in most other online communities (Nielsen, 2006). In addition, the regular accesses of other users, especially when crowd monitoring becomes relevant (closing of markets), indicated that other users would have been ready to participate if there was a need. However, it cannot finally be said what would have happened without this single user. Third, only a period of about five months was observed and within there was only one event of greater public interest. It was possible to demonstrate that fraudulent cases were found in the last elections on historical data, but the "monitoring" effect was only showed within the last periods of the 2017 Niedersachsen State Election, which is arguably important, but much less polarizing than the 2017 German Federal Election. The concept will need to prove itself in other polarizing events, such as the upcoming Federal Elections. Finally, the problem of a missing ground truth is still prevalent. The operators list of known cases was applied as a benchmark, but already the first 48 hours yielded further, yet unknown, cases. There was no case since the introduction of the tool that was detected by the algorithm and not by the crowd - and many cases detected by the crowd and not by the algorithm - but even still one cannot finally be sure, how many more fraudulent and manipulative cases there may be in the data. Some cases could not finally be concluded, e.g., when suspicious traders claimed to have traded "inexperienced".

The next step, which was already briefly discussed in the suggestion for a gamma cycle, will be to introduce a voting mechanism. This enables the crowd to decide whether the reported suspicious cases are actually fraud/manipulation or normal behavior – especially for these cases in which the market operator has no valid evidence or manipulative behavior is obvious but not punishable. It is also necessary to assess to which extent such an instrument is also a preventive measure, since its existence deters potential fraudsters. This is addressed in the following Section 7.6. It is also subject to further research, how the tool may be misused. This may be, e.g., to create wrong accusations, create spam reports, or to derive information on other users strategies in order to gain more profits. After all, such a tool can only work if the crowd engages in the tool in the long-term, which cannot be finally concluded yet.

The current work contributes to the academic research at several points. It provided an insightful case, in which the applicability of a crowd-based fraud and manipulation detection approach could be demonstrated. The theoretical derivation of the DPs and the firsthand experience contribute to the current academic research and practice by demonstrating a crowd-based approach to detect rare and ambiguous fraudulent and manipulative events in prediction markets. This theoretical derivation contained, in addition, the first (as far as known) collection of reported manipulative and fraudulent events and experiments in this area that summarizes the perceived effects on the market price. To conclude, a crowd-based approach to detect manipulative and fraudulent behavior is very promising for multiple applications in public and enterprise prediction markets.

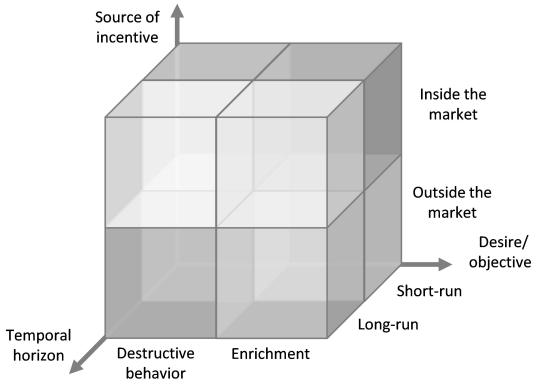


FIGURE 7.9: The Fraud Cube: Framework to understand and uncover where a prediction market may be manipulated or cheated.

7.6 From Detection to Prevention

7.6.1 The Fraud Cube: A motivational perspective on Manipulation and Fraud

Contents of this section are in part adopted from Kloker and Kranz (2017). See Section A.1 for further details.

To prevent manipulative behavior and fraud, it is important to understand, what and how fraudsters think, what they want, and what finally convinces them to manipulate and or break the prediction markets rules. Wolfe and Hermanson (2004) describe a fraudsters thought process in the "fraud diamond", which is an extension of the "fraud triangle" (Cressey, 1953; Lou and Wang, 2011). The fraud diamond states four necessary preconditions for fraud to occur: (1) the right person (capability) must realize (2) an opportunity (weakness in the system) and be (3) planning to do so (want or have to commit); finally this person has to be (4) convinced (rationalization) that the potential gains are worth the risk. In a prediction market, a willing person's intention can be described in three dimensions: (i) Desire/objective (whether to disrupt the market or to enrich itself), (ii) temporal horizon (immediately or in the long-run), and (iii) source of incentive (the issuing incentive is caused by an inner incentive scheme outside the market). To understand and organize motivational factors as well as attacking strategies better, the Fraud Cube (see Figure 7.9) organizes fraudulent patterns on three dimensions (i-iii). These are described in the following.

The desire of a fraudster (i) can be (self-) enrichment, destructive behavior, or both. In the case of (self-) enrichment the traders want to improve their own results (Bohm and Sonnegard, 1999) and realize that it is possible by playing against the rules. For "destructive behavior", there are mainly two reasons known, though there may be several others. On the one hand, according to Brüggelambert (1999), destructive participants take a delight in nonsensical decisions or are glad about sabotaging. On the other hand, they may have incentives for bad predictions. Such incentives may not only be monetary and may lay outside the market, such as in the 1999 Berlin State Election example mentioned earlier.

The temporal horizon of fraudulent behavior (ii) can be short- or long-run. Short--run attacking usually intends to realize quick profits or to destroy market prediction

TABLE 7.3: Using the Fraud Cube, illustrated at the DARPA Terrorism Futures example

Vector	**Motivation & Attack**
[Enrichment & Long-run & Outside]	A company/organization, selling products for security or earning money with the fear of an attack, may be given an incentive to write threatening letters to the government that claim that a terror strike will occur soon (Hanson, 2006a).
[Enrichment & Long-run & Inside]	If the prizes are very high (the inner incentive scheme is very strong), a trader may be incited to conduct a terror strike to make the "prediction" come true.
[Destruction & Long-run & Outside]	Terrorists may be encouraged to join the market and trade prices artificially down, in order to create a false sense for security. Or they may, at least, observe the prices to select a "good" date for their plan.

for only a short period. There is no distinct threshold which time period is needed for an attack to be considered short- or long-run. However, long-run fraudulent behavior takes a greater interest in the outcome of the event. They want to either manipulate the outcome or the information (and decisions) that are derived from the market prediction.

Finally, the incentive (iii) to cheat and manipulate may have two manifestations: It may come from inside or outside the markets. The examples mentioned earlier show both cases: While the manipulation of the election forecasts, e.g. 1999 Berlin State Election, have been incited from outside, the fraud that occurred in the FIFA World Cup example was incited from the incentive scheme of the market itself.

At this point, it needs to be mentioned that it is of great importance, thinking trough the different parts of the cube, to consider both: attacks from inside (trading) or outside (manipulate the information or outcome).

The DARPA Policy Analysis Market provides a good example, how to elaborate risks of fraud and manipulation in prediction markets by using the Fraud Cube, which is reported by Hanson (2006b) and Hanson (2006a). The market was discussed on July 2003, but quickly dropped for the reason of unpredictable risks. The project intended to create a prediction market to forecast assassinations and terror strikes. Comprehensibly, such a market has, besides of the regular inside incentive to perform well, strong outside incentives to be manipulated. Moving through the Fraud Cube, several attacking points can be identified, illustrated in Table 7.3:

Hanson (2006b) states that attacks from outside are easily forgotten when thinking about threats to the sustainable and correct operation of the market. The framework of the Fraud Cube helps market engineers to uncover attacking points by systematically thinking through the dimensions of fraud. However, due to the manifold expressions of attacking points and the fact that they are often not under the reach of a market engineer, it is hard to protect against them (Hanson, 2006b). Incentives may play a key role (Schröder, 2009), which will be discussed in Section 7.6.3.

7.6.2 How Fraud in Prediction Markets occurs

> Contents of this section are in part adopted from Kloker and Kranz (2017).
> See Section A.1 for further details.

To classify the attacks and manipulation strategies more in detail, categories from Allen and Gale (1992) are utilized, which originated in stock markets: (i) Action-based manipulation, (ii) information-based manipulation, and (iii) trade-based manipulation.

Action-based manipulation and information-based manipulation are not always easy to differentiate. In general, action-based manipulation is a manipulation of the value of the contract. Allen and Gale (1992) are drawing near the example of the American Steel and Wire Company from 1901. The managers of the company shortened their stock positions and then closed the steel mills. The price fall from $60 to $40. The managers covered their short positions and reopened the mills, which led to a price rise and to a large profit of the managers. Bagnoli and Lipman (1996) developed another model of action-based manipulations. Here a participant in the market pools with an external partner who is giving a takeover bid for the company the manipulator is holding stocks. After the manipulator has realized his profits, the partner unwinds his takeover bid. A similar model was presented by Vila (1989) besides that the roles are switched, so that the initiator is the giver of the takeover bid. Prediction markets are a common instrument to predict project milestones or product release dates. Hanson (2006b) states here the problem that employees, trading on the market, may take influence on the outcome. Ottaviani and Sørensen (2007) showed in an experiment that employees had indeed taken the opportunity, if given so, to manipulate the outcome in such a context. Chakraborty and Das (2016) investigate this problem from a game-theoretic perspective and conclude also that the opportunity lead to manipulation or fraud.

Information-based manipulation is related to the spreading of false information or of deceptive rumors. Examples for this kind of manipulation are the "trading pools" that emerged in 1920 in the US, the Enron, and the WorldCom frauds in 2001 (Unerman and O'Dwyer, 2004). According to Benabou and Laroque (1992) an opportunistic trader with privileged information could profitably manipulate markets by making misleading announcements in case he seems credible to other investors. Recently, Casas, Fawaz, and Trindade (2016) demonstrated this effect in prediction markets in the context of elections.

Trade-based manipulation is manipulation playing within the rules. This is especially researched in stock markets (Allen and Gale, 1992). The basic consent is that trade-based manipulation is possible given some preconditions (usually true in prediction markets: low liquidity, non-linear demand functions). However, profits are low and if such behavior in play-money prediction markets lead to higher engagement it is often not considered as fraudulent behavior or harmful (e.g. Hanson and Oprea, 2009).

In the context of play-money prediction markets, these three types of manipulation and fraud have to be extended by one more type: (iv) Multiple accounts and coalitions. Blume, Luckner, and Weinhardt (2010) and Blume (2012) presented several strategies that are used by participants which have created (or hacked) multiple accounts. This is usually done in order to trade between these accounts and transfer money from one to the other (Schröder, 2009). Similar problems can arise if participants start to form coalitions. This may be the case if participants are given the incentive of prizes for the top ranks. If the prizes can be shared among the coalition partners, they are able to increase the probability over the average probability

of winning the price, if they transfer money (or stocks) to one account. This is possible, if the spread is greater or equal to 0.03 MU. This strategy to transfer money is called "ping-pong" or circular trading, which is also described in Hansen, Schmidt, and Strobel (2004) and even in real stock markets (Reuters, 2010). The strategy is more successful, the higher the spread is, so sometimes the spread is aggressively widened before the circle trading (Blume, Luckner, and Weinhardt, 2010).

7.6.3 Prevention and the Role of Incentives

> Contents of this section are in part adopted from Kloker and Kranz (2017).
> See Section A.1 for further details.

The Fraud Cube, introduced in Section 7.6.1, already shows that it is not enough to only detect fraudulent actions by searching for fraudulent patterns in the trading. Fraud prevention starts much earlier. According to the "fraud diamond" (Wolfe and Hermanson, 2004) fraud (and manipulation) has four preconditions. Against the first, the "capability", unfortunately there are no countermeasures.

What about the "weakness in the system"? Especially in play-money prediction markets, but basically in all prediction markets, problems arise if traders form coalitions or if one trader controls more than one account (no matter if he created or hacked them). These weaknesses cannot be turned off, but several countermeasures can hamper this kind of fraud. In the examples of Hansen, Schmidt, and Strobel (2004) it is described that a delay in activating new accounts efficiently helped to slow down and reduce the effect of fraud. This may already scare off fraudsters that are only interested in "quick returns". However, there remains the trade-off between security and comfort of registration. Huang (2016) and Huang and Shoham (2014) suggest to implement trading limits in the context of prediction markets featuring a market scoring rule. This idea transferred to markets where supply and demand meets directly would mean to limit the amount of stocks (or money) which can be traded between two accounts. There may also be a time component in this restriction. Though this may effectively slow down fraud by hampering common trading strategies, there is also a trade-off: Traders may perceive the match of trades not comprehensible when their trades are mapped to "suboptimal" orders (as the "optimal match was prevented by volume restrictions). Other measures to decrease fraud opportunities is a careful contract design (Hansen, Schmidt, and Strobel, 2004) or the deployment of bots (George Mason University, 2015). As a countermeasure against politically motivated manipulation of prices, Berlemann and Schmidt (2001) suggested to show only long averages on the prices as forecasts (24h). As it is hard to manipulate prices for long periods, this may discourage manipulators, as their actions have only marginal effect on the forecasts.

Probably the most promising approach to prevent fraud is thinking about the role of incentives. Incentive compatibility means that for a risk-neutral agent the best strategy is to follow the rules. This may be reached if all participants truthfully reveal any private information asked for by the market mechanism (Schröder, 2009). A market is not incentive compatible if traders can obtain a higher profit through fraudulent actions or if participants avoid to participate as they fear manipulation. A good, although admittedly hard, way to prevent manipulative behavior within the market, is to create a "[...] design that makes it more lucrative to play according to the rules" (Schröder, 2009, p. 18). For manipulative behavior to happen in prediction markets, there must be an incentive inside or outside the market that is stronger than the incentive that can be reached by playing according to the rules. Therefore,

it is important to think about which conflicting incentives can occur for traders or groups.

Dimitrov and Sami (2010) is presenting a setting of two parallel prediction markets with different incentive schemes. Participants have the ability to trade on both markets. If the incentive in the second market is stronger, manipulative agents can start to trade irrational in the first market, to make the price in the second market follow and to open profit options. Dimitrov and Sami (2010) characterized the weak perfect Bayesian equilibrium for such a case and found that the payoffs were unique across all equilibria. Chen and Kash (2011) consider conflicting incentives in prediction markets with scoring rules that are used for decision-making. They argue that the experts trading in the market do not necessarily want to select the optimal decision, but generate the highest payoff. If the scoring rule is not suitable for truthful elicitation of information, the incentive for the expert is in conflict with the goal of a prediction market operator. The problem in general, when prediction markets are used for decision-making, is examined by Othman and Sandholm (2010). For some participants this can create strong outside incentives, if they would benefit of certain outcomes. They can be incited to strategically manipulate the market probability. If this issue is known in a market, it can also lead to distrust between the participants. Chen et al. (2015) studied conflicting incentives inside and outside the market. They concluded that conflicting incentives not necessarily may damage information aggregation in equilibrium and showed some examples. However, there were also many situations, where information loss was inevitable. In addition, Malekovic, Sutanto, and Goutas (2016) found that the intention to manipulate in a setting comparable to a prediction market using a market scoring rule, raises with the information asymmetry between the participants. It also decreases with the complexity of the scoring rule.

A prediction market engineer should keep conflicting incentives in mind when designing the market and contracts. This should include considering if there are other prediction markets that offer comparable contracts with different incentive schemes, and adapt accordingly or even drop the own contract. It has also to be considered how results are communicated, which persons should be excluded from participation and to which extent decisions are directly bound to the outcome of prediction markets. In markets where prizes are raffled among the best traders, this may include to only offer "non-sharable" prizes (in contrast to money, as in Bohm and Sonnegard (1999)). As the price may not be shared among a coalition, the incentive for the coalition to work together becomes weaker. To sum up, fraud to occur needs incentives and thinking about conflicting incentives in advance may reduce fraud and manipulation distinctly.

Finally, according to the "fraud diamond", a trader has to be convinced, rationalized that the fraud and manipulation is worth the risk. Several small design adaptions are conceivable in prediction markets that may help manipulators and or fraudsters to change their mind. Some of them have been suggested in various academic literature. A brief summary can be found in Table 7.4. However, no empirical evidence on such a measure is reported in the literature yet.

TABLE 7.4: List of suggested or discussed design elements for the prevention of manipulation

Reference(s)	**Suggestion/Findings**
Publicity and Access	

TABLE 7.4: List of suggested or discussed design elements for the prevention of manipulation

Reference(s)	**Suggestion/Findings**
Rhode and Strumpf (2008)	Non-public markets: No differences between anonymous and public markets regarding manipulation.
Hansen, Schmidt, and Strobel (2004) and Bohm and Sonnegard (1999)	Reduce press coverage.
Hansen, Schmidt, and Strobel (2004)	Publication delay in press coverage hinders manipulation.
Hansen, Schmidt, and Strobel (2004)	Account activation delay hinders manipulation.
Hansen, Schmidt, and Strobel (2004)	Display long averages as predictions hinders manipulation.
Ho and Chen (2007) from Buckley and Doyle (2017)	"Limit the Stake" limits manipulation.
George Mason University (2015)	Do not use large extrinsic prizes
George Mason University (2015)	Require more user information
George Mason University (2015)	Publicize blocking of users
Contract Design	
Hansen, Schmidt, and Strobel (2004)	Trade small parties within the "other parties" stock or trade them as winner-takes-all contracts. Second suggestion is, however, unclear to which extent this would solve the problem. There is more capital needed to change the decision in case of clear losers, but not in case of high uncertainty.
Huang and Shoham (2014)	Implement trading limits (in markets with a Market Scoring Rule).
This work (Section 7.6.4)	Opening auction to prevent shallow markets in the beginning.
Technical Design	
George Mason University (2015)	Bots make collusion more difficult. If the set up for bots is facilitated for more users, this may prevent fraud.
Kloker and Kranz (2017)	Use warnings based on suspicious points.
This work (Section 7.6.4)	Use categories for the ranking to ensure that traders, knowledgeable in only one category, find their expertise reflected.

7.6.4 Adaptions to the FAZ.NET-Orakel

Contents of this section are in part adopted from Kloker et al. (2018b).
See Section A.1 for further details.

Warnings

Initial cases of fraud in the FAZ.NET-Orakel and feedback from single fraudsters showed that the participants were obviously not aware of the detection algorithm or felt smarter than the algorithm. Some even reported that they were not aware that manipulative or fraudulent attempts may lead to an exclusion from the market (though this was written in the terms). The George Mason University (2015) reported that after the publication of blocked accounts, manipulation was virtually not existent. This suggests that it may help to prevent manipulation and fraud if traders are made aware of the consequences. For this reason warnings were displayed to traders, when their trading behavior showed a hint of fraudulent patterns. If the pattern continued their account was locked, displaying a message that they should contact the administrator. A screen-shot of the later message is illustrated in Figure 7.10.

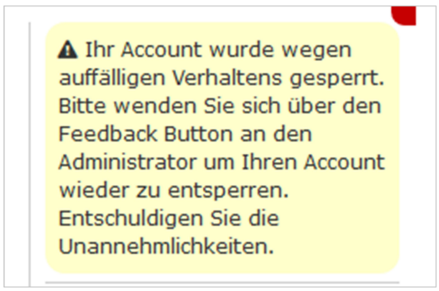


FIGURE 7.10: Message "Account locked" as displayed to traders that showed a suspicious pattern.

Topic-specific Rankings

In May 2017 a fraudulent account agreed to participate in a telephone interview. This revealed a very interesting and yet not considered source of incentive in literature. The trader was especially knowledgeable in areas of economics. Indeed, he won the tournament of the last EIX version, which focused mainly on economic indicators. The focus of the FAZ.NET-Orakel is, however, much broader, why his good performances in the economic indicators was opposed to a moderate performance in other topics (politics or entertainment). This was diametrical to his self-image and frustrated him, for which reason he started to perform fraud in non-economic markets until he reached a position in the ranking that he perceived to be fair. To prevent or at least hamper this kind of fraud, topic specific rankings were introduced alongside the already existing overall ranking (see Figure 7.11). Hence, participation in all markets is not mandatory in order to be listed in the top positions of the rankings regarding certain topics.

Opening Auctions

In August 2017 several users complained that especially at the market start, low liquidity is used to gain windfall profits from uninformed traders. Bohm and Sonnegard (1999) already found shallow markets to be a popular aim for fraudulent behavior, as well. To prevent such wind-fall profits, opening auctions were introduced. Opening auctions are starting 24h before the normal CDA trading is possible. All submitted orders within this time are collected and centrally matched at the

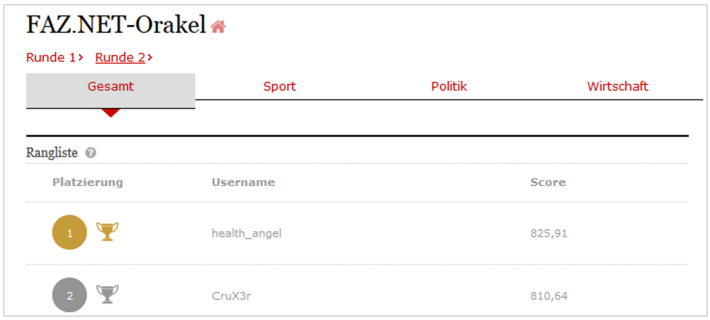


FIGURE 7.11: Screen-shot of the ranking overview. Several sub rankings were available.

FIGURE 7.12: Screen-shot of the trading screen during the opening auction.

end of the opening auction phase at the price that results in the maximum transfer of stocks. By doing so, it is ensured that there is a certain liquidity at the beginning of the market and the start price is fair. Hence, there is no shallow market and potential windfall profits are equally distributed among all informed traders. The trading interface during the opening auction is illustrated in Figure 7.12.

Tool for Crowd-based Manipulation and Fraud Detection

In October 2017 a survey revealed that still many users perceived that there exists cheating on the market. Potentially by patterns that are too creative for the current algorithm. Some users offered to support detecting these fraudulent accounts, if they got access to the trading data. For this reason the tool for crowd-based manipulation and fraud detection was developed that was described in more detail in Section 7.4. However, besides the detection of fraudulent and manipulative actions, it has some preventive character as well. The tool is accessible from each market and has therefore some visibility (though not overly emphasized).

7.6.5 Evaluation of the Adaptions and Conclusion

> Contents of this section are in part adopted from Kloker et al. (2018b) and Kloker et al. (N.D.).
> See Section A.1 for further details.

To evaluate the applicability of countermeasures against manipulation and fraud is always very hard, as the phenomenon cannot be reasonably induced in the FAZ-.NET-Orakel setting and its occurrence is also dependent on many other variables (overall activity, markets, topics, etc.). Assess preventive design elements is even harder. Therefore, the features were evaluated continuously in close exchange and discussion with the super-users of the FAZ.NET-Orakel. Briefly, it can be concluded that after the warnings and the updated ranking virtually no fraudulent action was observed anymore, besides by some intentionally losing accounts (as also observed by the George Mason University (2015)). Manipulation was, however, still persistent. After the introduction of the opening auction, complaints regarding fraud during low liquidity phases was decreased. To get an impression on the preventive character of the crowd-based manipulation and fraud detection tool, a survey among all active participants (more than 10 orders placed, N = 514) was performed, of which n = 29 responded. Among others, the participants were asked how easy they perceived the opportunity for manipulation and fraud before and after the introduction. Although the sample size was small (n = 29), a one-tailed, paired Wilcoxon Rank Sum test showed that the perceived opportunity was significant lower after the introduction of the artifact (p = .087).

However, especially regarding manipulation, during the first year there was only one event where manipulation was to be expected, as it was politically controversial and the incentives from outside of the market were high (Deck, Lin, and Porter, 2013; Rhode and Strumpf, 2008). Hence, it is hard to separate which underlying reason, the implementations or the missing incentives, led to virtually no observations of manipulation.

Still, the adaptions can be regarded as successful and can be suggested for implementation in future prediction markets as this relatively small effort may have a huge impact and distinctly raise prediction quality and participation.

Chapter 8

Finale

Have a message and be one.

Oswald Chambers

8.1 Conclusion

This thesis instantiated one Delphi-Market in the form of the FAZ.NET-Orakel. In Section 3, the diversity of potential integration approaches already suggests that it is not possible to evaluate the concept of Delphi-Markets with only one artifact. Instead, besides elaborating the concept theoretically, the instantiated artifact was used to improve the crowd-based judgmental forecasting methodology in four ways. These four research projects were motivated by the JFIM that was introduced in Section 2.1.3 and which corresponds to the four dimensions of error by Armstrong (1985). In the JFIM these errors are located at the intersections of the six core aspects of forecasting methods by Lyon and Pacuit (2013) and the factors "motivation" and "cognition", which enrich the six core aspects with a human component.

With the first research project outlined in Section 4, it was demonstrated that prediction market trading behavior is an indicator to select potentially knowledgeable participants as experts for RTD studies. The sampling is, however, not only limited to the selection for RTDs. Companies applying prediction markets may identify knowledgeable employees for other forms of expert panels or to better understand the importance of an individual employee in a team for its information flow. The research project showed the general applicability of the approach to select informed traders or high-performers based on historical data. To finally assess the accuracy and information gain induced by this approach, a long-term field study would be required. Theory, however, suggests that this selection approach may improve the diversity of viewpoints and information considered (Welty, 1972; Green, Armstrong, and Graefe, 2007) as well as the general rigor of the Delphi methodology (Hasson and Keeney, 2011). In the FAZ.NET-Orakel, anecdotal evidence showed that self-selection of participants on the RTD that were recruited among the prediction market participants leads to a very similar panel. Mostly those participants took part in the RTD that also performed significantly well in the prediction market. It is subject to future research how the information flows are designed in detail and to distinguish if high-performers are likely to take part in RTD studies or if the participation in RTD studies leads to better prediction market performance. In Chapter 4 only the suggested selection approach "The Potential" (see Section 3.1.2) was implemented. Future work may evaluate the other approaches, especially "The Bohemian" as a potentially interesting approach to further enhance expert selection.

In the second research project outlined in Section 5, a social RTD approach was motivated, suggested, and evaluated in order to address the problem of low retention in Delphi studies. This sRTD was implemented in the FAZ.NET-Orakel. In a preliminary online experiment and a subsequent field study in the context of the 2017 German Federal Election it could be demonstrated that allowing social interaction in form of labeling in the discussion can raise retention among the participants over Delphi rounds. The effect in the field study seemed to be explained by the positive labels. As this was a field experiment, not only the intention to participate, but also the actual behavior could be measured. For this reason it is to assume that the approach would work in other real-world settings as well. Allowing social interaction, though there is full anonymity, raises retention. This may be especially interesting in company settings, where a workers council or a labor union are very sensitive towards every form of participation that may potentially allow the identification of individual employees. Due to restrictions and low general participation in the RTD these results should, however, be verified by future long-term studies. Yet, it also cannot be said what effect the certain topical background had on the results. Interactions between the prediction market component and the RTD seemed to have no effect on the retention. Therefore, a cautious conclusion may also be that the theoretical potential derived in Section 3 to use prediction markets in order to provide incentives for the RTD may not hold at least in the integration approach on a user-level.

In the third research project outlined in Section 6, a common problem of many forecasting methods, but especially prediction markets, is addressed: The partition dependence bias. In two consecutive online experiments the moderators "complexity" and "expertise" were elaborated. Both showed an effect on the occurrence of partition dependence. However, though complexity moderated the partition dependence, it was yet not possible to finally assess in what way it is related to a raising or decreasing effect. The reason for this is that measuring the cognitive load and therefore active processing type according to the Dual-Process Theory turned out to be unreliable. Expertise, however, systematically decreased partition dependence. In the context of Delphi-Markets it therefore has to be considered, if this may be an additional potential for an integrated approach. Prokesch, Gracht, and Wohlenberg (2015) already selected participants for a hidden market. The experiments in Section 6, in addition, reproduced the effect previously found by Sonnemann et al. (2013) in CDA markets for LMSR markets. This emphasizes the importance to further understand the phenomenon of partition dependence. While for expertise a decreasing effect was demonstrated, future research should address a more in-depth elaboration of complexity and further moderators such as motivation.

In the fourth research project outlined in Section 7, the FAZ.NET-Orakel was extended by a crowd-based manipulation and fraud detection artifact (and several further small design improvements) in order to address the problem of manipulation and fraud. Manipulation and fraud is one of the five open questions regarding prediction markets (Wolfers and Zitzewitz, 2006a; Buckley and Doyle, 2017). The suggested and implemented artifact addresses several yet unsolved problems: (1) The detection of (creative) attacking strategies beyond the capabilities of rule-based algorithms. (2) The detection of manipulative patterns that may harm the market but are not necessarily against the rules. (3) A mechanism to deal with inconclusive cases of manipulation and fraud for those cases in that the operator would not have any means to deal with. In addition, the artifact may have a preventive character and, therefore, improve the market quality even before price manipulations or

fraud happen. These findings and the underlying approach are of relevance for every setting, where prediction markets are only run as a supplementary instrument and should not require too much attention of the IT department for supervision and development. The approach is not based on complex algorithms and may even work entirely without any administrator to ultimately decide on manipulative and fraudulent cases. The approach enabled the crowd to discuss on ambiguous cases and anecdotal evidence showed that the discussions on the few suspicious cases were carried out on a content level. Participants tried to comprehend the potentially underlying information. Therefore, the introduction of this tool may even have the same positive effects as attributed to the integration on a market-level approach (see Section 3.1.2), though to a very limited scope. Yet, there is not sufficient long-term data in order to evaluate long-range effects of this tool on the market. In addition, it cannot yet be concluded how and to which extent the tool may be misused or subject to manipulation and fraud itself. However, it provides a simple and convenient way that is easy to implement for practitioners, to secure a prediction market against a wide spectrum of potential attacks beyond the capabilities of detection algorithms[1].

The current thesis contributes with four research projects in the context of the FAZ.NET-Orakel to four prevalent problems in judgmental forecasting methods (especially prediction markets). In addition, it is the first implementation and elaboration of a Delphi-Market with public access that implemented a user-based integration. Although many of the suggested design elements and artifacts may need a further validation with long-term data, they may help researchers and practitioners in building better judgmental forecasting tools and understand common errors as well as their reduction.

8.2 Outlook

Future of current research

Finally, the question remains, how the future of the Delphi-Markets will look like? Section 3 outlined that there are plenty of possible approaches to integrate prediction markets and RTD. Yet, it is not clear if there is "the one" dominant design. Probably, the design decision should always be made in light of the context of each specific forecasting task and potential challenges. The FAZ.NET-Orakel will be continued in order to collect more long-term data and to re-validate the previous findings, both regarding the retention in RTD and the crowd-based manipulation and fraud detection tool. It is of special interest to observe these two artifacts in the context of a forecasting event of high interest, such as the upcoming German Federal Elections. In addition, the expert selection algorithm elaborated and suggested in Section 4, which was evaluated on historical data, also has to prove itself on "live" data. This, even though already implemented in the FAZ.NET-Orakel, remains as an outstanding task.

Research on the partition dependence bias should focus on finding experimental designs that allow to decompose the effect of individual moderators. The experiments outlined in Section 6 can be understood as an exploration of possible moderators, which involved more or less rudimentary complexity, expertise, motivation, risk affinity, and enjoyment. While complexity and expertise seemed to have the most explaining power, more isolated evaluation of each individual moderator is required in order to fully understand the bias.

[1] Though it is not argued that the tool may not be run accompanying a detection algorithm.

Future research opportunities

Prediction markets as a way of participation In a broader perspective, prediction markets will probably gain in importance in companies and in politics. Participation in companies is a growing field that currently receives a lot of attention in academic research (Wagenknecht, Filpe, and Weinhardt, 2017; Niemeyer et al., 2016). Prediction markets have become more and more adaptable and easy to use and, therefore, qualify for decision support and forecasting in companies (Buckley, 2016). They provide a convenient way to allow employees express their estimation on questions regarding the future, new ideas, or even the success of managerial measures. However, in which way prediction markets and Delphi-Markets influence the willingness of employees to participate, the quality of decision-making, and the overall culture in a company is yet an open research question.

Predictions based on distorted samples In politics prediction market feature some favorable properties, especially regarding current trends in the political discourse, such as greater polarization and stigmatization, and the extended use of social networks (Yang et al., 2016; Brug, Fennema, and Tillie, 2000; Lee et al., 2014). Prediction markets can adapt to fast changing information better than many other methods. In addition, they are relatively robust against distorted samples and fear of stigmatization as participants remain anonymous. And though participants do not have to reveal their preferences, the trading behavior can be used as a kind of Bayesian Truth Serum, as it was shown that the depot structure of individual traders can be related to party preferences with high accuracy (Kranz et al., 2014). In order to foster these favorable properties, prediction markets may be integrated with online social networks. Such an integration may help to reduce the judgment bias in the predictions (Graefe, 2014; Kranz et al., 2014). An integration of a prediction market with a social network was also already demonstrated recently by Qiu and Kumar (2017).

From expectation to opinion With an increasing reach of prediction markets, especially towards participants that are not used to markets in general, the interfaces of prediction markets have become more easy (hidden markets) and intuitive. Such a simple market interface is implemented, for instance, in the MicroMarkets interface (see Section 3.2.3). These almost gamified interfaces, however, raise the question to which extent they really still collect the expectation of a participant or only a mere opinion. In addition, the playful design of the MicroMarket may stimulate risk-taking (nudged to move the sliders more towards the ends). These factors put the challenge to the researcher, how the individual inputs should be interpreted and potentially modified before aggregated in a market mechanisms – or if this is not necessary at all.

Hybrid markets It is also very likely that prediction markets will be populated not only with human, but also artificial agents in future (regardless if this "bots" act independently or with instructions and in the name of real persons). Such hybrid markets have already been tested and evaluated positively (Nagar and Malone, 2012) and the George Mason University (2015) suggested them as a measure against manipulation and fraud. However, which potentials hybrid markets bear is yet an open research question.

Appendix A

Appendix

A.1 Disclosure of own contributions

It is common in academic research that in the continuous discussion and exchange with other researchers a research project is improved, shaped, evaluated, and even sometimes completely inverted. Research is to a large extent team work, as in many cases a single person would not be able to perform the data collection alone or know all available literature by heart. Same is true for presented research, many hands and heads shaped the results to small or sometimes even significant extents. These persons were attributed an appropriate credit, sometimes also resulting in a co-authorship of those papers on which this thesis is based. This section intends to constitute in detail which of the parts were performed by the author of this thesis and which parts were a joint work, in order to help the reader assess the efforts and achievements of the author's work.

Kloker et al. (2016) is a joint paper with Dr. Tobias T. Kranz, Dr. Tim Straub, and Prof. Dr. Christof Weinhardt, published as a Full Paper in the Proceedings of the Second Karlsruhe Service Summit Research Workshop 2016. My contributions consisted of:

- The literature review.
- The theoretical foundations to suggested approaches.
- The formulation of the proposal.
- The writing of all sections.
- The presentation of the paper.

Kloker (2016) is a single-authored working paper.

Kloker, Straub, and Weinhardt (2017a) is a joint paper with Dr. Tim Straub and Prof. Dr. Christof Weinhardt, published as a Prototype Paper at DESRIST 2017. My contributions consisted of:

- The literature review.
- The refinement of the research question.
- The suggested selection strategies ii) and iii) and their derivation.
- The artifact development.
- The writing of all sections.
- The presentation of the paper.

Kloker and Kranz (2017) is a joint paper with Dr. Tobias T. Kranz, published as a Research in Progress Paper in the Proceedings of the 25th European Conference of Information Systems 2017. My contributions consisted of:

- The literature review regarding attacking strategies and sources of incentives.
- The extension of the fraud cube with the dimension "source of incentive".
- The implementation and calculation of the benchmark scenario.
- The comparison of the suggested approach with the benchmark scenario.
- The discussion.
- The writing of all sections.
- The presentation of the paper.

Kloker et al. (2017) is a joint paper with Dr. Tim Straub, Dr. Tom Zentek, and Prof. Dr. Christof Weinhardt, published as a Research in Progress Paper in the Proceedings of the 17th International Conference on Group Decision and Negotiation 2017. My contributions consisted of:

- The literature review.
- The derivation and formulation of the research question.
- The draft for the research design.
- The draft for the experiment design.
- The implementation and execution of the experiment.
- The implementation of the evaluation scripts.
- The evaluation with several consultation cycles.
- The discussion.
- The writing of all sections.

Kloker, Straub, and Weinhardt (2017b) is a joint paper with Dr. Tim Straub, and Prof. Dr. Christof Weinhardt, accepted and presented at the Collective Intelligence Conference 2017. My contributions consisted of:

- The literature review.
- The derivation and formulation of the research question.
- The draft for the research design.
- The draft for the experiment design.
- The implementation and execution of the experiment.
- The implementation of the evaluation scripts.
- The evaluation with several consultation cycles.
- The discussion.
- The writing of all sections.

Kloker et al. (2018c) is a joint paper with Dr. Tim Straub, Dr. Stefan Morana, and Prof. Dr. Christof Weinhardt, published as a Full Paper in the Proceedings of the 26th European Conference on Information Systems 2018. My contributions consisted of:

- The literature review.
- The derivation and refinement of the research question based on Kloker et al. (2016).
- The draft for the research design.
- The experiment design for the online experiment based on Kloker et al. (2016).
- The implementation and execution of the online experiment.
- The implementation of the evaluation scripts.
- The evaluation with several consultation cycles.
- The study design for the field study.
- The implementation and execution of the field study.
- The evaluation and discussion.
- The writing of all sections.
- The presentation of the paper.

Kloker et al. (2018b) is a joint paper with Dr. Tim Straub, Dr. Stefan Morana, and Prof. Dr. Christof Weinhardt, presented as a Prototype Paper at DESRIST 2018. My contributions consisted of:

- The literature review.
- The derivation and formulation of the research question.
- The design and implementation of the design elements (1-3).
- The design and implementation of the crowd-based manipulation and detection tool (4).
- The survey design of the evaluation survey.
- The implementation and execution of the online survey.
- The evaluation of the online survey.
- The writing of all sections.

Kloker et al. (2018a) is a joint paper with Frederik Klatt, Jan Höffer and Prof. Dr. Christof Weinhardt, published as a Full Paper in the "Foresight" journal. My contributions consisted of:

- The derivation and formulation of the research question.
- Guidance regarding the research method.
- Guidance regarding the literature review.
- Guidance regarding the research design.
- The data and basic processing.
- Guidance and help regarding the feature selection, implementation, and evaluation.
- Guidance regarding the interpretation of the results.
- The discussion.
- The writing of all sections.

Kloker et al. (N.D.) is a joint paper with Dr. Tim Straub, Dr. Stefan Morana, and Prof. Dr. Christof Weinhardt, currently unpublished and in the state of a Working Paper. My contributions consisted of:

- The literature review.
- The derivation and formulation of the research question.
- The design and implementation of the crowd-based manipulation and detection tool.
- The development, operation and maintenance of the FAZ.NET-Orakel.
- The project management and all contacts to the project partners and super-users.
- The survey design of the evaluation survey.
- The implementation and execution of the online survey.
- The evaluation of the online survey.
- The discussion.
- The writing of all sections.

Kloker et al. (2019) is a joint paper with Dr. Tim Straub, Dr. Tobias T. Kranz, and Prof. Dr. Christof Weinhardt, accepted in an reduced Version to be published as a Chapter in "Delphi-Verfahren: Konzept, Varianten und Einsatzbereiche in der Gesundheitswissenschaft", eds. M. Niederberger, O. Renn. My contributions consisted of:

- The literature review.
- The co-derivation of the potentials.
- The derivation of the challenges.
- The development of the User-level and Delphi-Question-level approaches.
- The formalization and discussion of all approaches.
- The writing of all sections.

Kloker, Straub, and Weinhardt (N.D.) is a joint paper with Dr. Tim Straub and Prof. Dr. Christof Weinhardt, under review in an extended version as a Full Paper in the "Group Decision and Negotiation Journal". My contributions consisted of:

- The literature review.
- The derivation and formulation of the research questions based on Kloker (2016) and Kloker et al. (2017).
- The section on experiment 1 is adapted from Kloker et al. (2017).
- The experiment design for experiment 2.
- The implementation and execution of experiment 2.
- The evaluation and discussion of experiment 2.
- The general discussion.
- The writing of all sections.

A.2 Different trees for trading-based expert selection

TABLE A.1: Different trees for trading-based expert selection (some examples). The first column indicates if only the identified three (market maker, spread, trading volume) or all attributes were provided to the tree algorithm. The attributes listed in the second column show, which attributes were considered by the tree.

	Atttributes	minsplit	Accuracy	Precision
three	Market maker	40	85%	47%
three	Market maker, Trading volume	20	88%	71%
three	Market maker, Trading volume, Spread	17	87%	60%
all	Market maker, Trading volume, Spread, OB size	15	86%	54%
three	Market maker, Trading volume, Spread	15	87%	60%
all	Market maker, Trading volume, Spread, OB size, market activity, limit order	14	86%	54%
three	Market maker, Trading volume, Spread	14	87%	60%
all	Market maker, Trading volume, Spread, OB size, market activity, limit order	13	88%	64%
three	Market maker, Trading volume, Spread	13	88%	75%
all	Market maker, Trading volume, Spread, OB size, market activity, limit order	12	89%	69%
three	Market maker, Trading volume, Spread	12	88%	75%
three	Market maker, Trading volume, Spread	10	86%	54,5%

A.3 Questionnaire items

TABLE A.2: Survey items of the follow up questionnaire of the preliminary experiment in Section 5.6.1.

Item	**Subconstruct**	**Loading**	**Short**
Addressability			
Es war mir möglich ein Bild von manchen Teilnehmern des Surveys zu machen.	Social Comfort	positive	add01
Das Interface vermittelt mir das Gefühl, dass die anderen Teilnehmer und ich in einem persönlichen Bezug zueinander stehen.	Perceived Social Presence	positive	add02
Das Interface vermittelt mir das Gefühl, dass ich mit den anderen Menschen interagiere.	Perceived Social Presence	positive	add03
Das Interface vermittelt mir ein Gefühl von Geselligkeit.	Perceived Social Presence	positive	add04
Das Interface vermittelt mir das Gefühl, dass die anderen Teilnehmer menschlich sind.	Perceived Social Presence	positive	add05
Ich habe mich während des Surveys unwohl dabei gefühlt, meine Meinung auszudrücken.	Social Comfort	negative	add06
Das Interface gibt mir ein Gefühl der Zusammenarbeit mit den anderen Teilnehmern.	Spirit	positive	add07
Ich fühle mich wichtig in diesem Interface.	Spirit	positive	add08
Das Interface vermittelt mir kein Gemeinschaftsgefühl.	Spirit	negative	add09
Ich fühle mich in dem Interface fehl am Platz.	Spirit	negative	add10
Anonymity			
Ich hatte das Gefühl, dass die anderen Teilnehmer in der Lage waren, meine Antworten und Ideen zu mir zurückzuverfolgen.		negative	ano01
Commitment			
Ich bin dazu bereit mehr als normalerweise erwartet zu investieren.		positive	com01
Ich habe ein Gefühl von Loyalität den anderen Teilnehmern gegenüber.		positive	com02
Ich habe vor, den Survey weiter zu verfolgen.		positive	com03
Reputation			
Wenn ich mein Wissen mit den anderen Teilnehmern teile, erwarte ich, dass die anderen Teilnehmer ihr Wissen auch teilen.	Reciprocity	positive	rep01

TABLE A.2: Survey items of the follow up questionnaire of the preliminary experiment in Section 5.6.1.

Item	**Subconstruct**	**Loading**	**Short**
Es hat für mich keinen Vorteil mein Wissen mit den anderen Teilnehmern zu teilen.	Perceived Value	negative	rep02
Ich denke, die anderen Teilnehmer honorieren meinen Beitrag.	Feeling as an Expert	positive	rep03
Ich würde mich selbst als Experte in dem Thema bezeichnen.	Feeling as an Expert	positive	rep04
Ich hatte das Gefühl, dass mein Wissen keinen Mehrwert bringen konnte.	Feeling as an Expert	negative	rep05
Ich hatte das Gefühl, dass ich durch mein fachliches Wissen über das Thema einen wertvollen Beitrag beisteuern konnte.	Feeling as an Expert	positive	rep06
Ich denke, die anderen Teilnehmer profitieren von meinem Wissen.	Feeling as an Expert	positive	rep07
Es ist vorteilhaft für mich, mein Wissen mit den anderen Teilnehmern zu teilen.	Perceived Value	positive	rep08
Insgesamt hat Teilen mit den anderen Teilnehmern einen guten Gegenwert.	Perceived Value	positive	rep09

TABLE A.3: Items for knowledge and experience in the preliminary experiment for partition dependence in Section 6.3.

Item	**Construct**	**Loading**	**Coding**
Diesel Price			
Are you regularly drive a car?	Experience	positive	binary
Have you been refueling a car within the last 2 weeks?	Knowledge	positive	binary
DAX-30			
Are you interested in the DAX-30 in general?	Experience	positive	binary
Have you followed the development of the DAX-30 within the last two weeks?	Knowledge	positive	binary
Deutsche Bank' stock			
Are you interested in the Deutsche Bank' stock in general?	Experience	positive	binary
Have you followed the development of the Deutsche Bank' stock within the last two weeks?	Knowledge	positive	binary

TABLE A.4: Items for REI-10 in experiment 2 for partition dependence in Section 6.4.

Item	Construct	Loading	Coding
Ich mag es nicht, viel nachdenken zu müssen.	Need for Cognition	negative	Likert-5
Ich versuche Situationen zu vermeiden, die tiefgründiges Nachdenken über etwas erfordern.	Need for Cognition	negative	Likert-5
Ich bevorzuge etwas zu tun, das meine Denkfähigkeiten herausfordert gegenüber etwas, das wenig Nachdenken erfordert.	Need for Cognition	positive	Likert-5
Ich bevorzuge komplexe Probleme gegenüber einfachen Problemen.	Need for Cognition	positive	Likert-5
Es stellt mich kaum zufrieden, für lange Zeit über etwas scharf nachzudenken.	Need for Cognition	positive	Likert-5
Ich vertraue meinen anfänglichen Empfindungen gegenüber Menschen.	Faith in Intuition	positive	Likert-5
Ich glaube, dass ich meiner Intuition vertrauen kann.	Faith in Intuition	positive	Likert-5
Meine anfänglichen Eindrücke von Menschen sind fast immer richtig.	Faith in Intuition	positive	Likert-5
Wenn es darauf ankommt Menschen zu vertrauen, kann ich mich normalerweise auf mein Bauchgefühl verlassen.	Faith in Intuition	positive	Likert-5
Ich kann normalerweise spüren, ob eine Person richtig oder falsch liegt, auch wenn ich nicht erklären kann wie.	Faith in Intuition	positive	Likert-5

TABLE A.5: Items for the context specific questionnaire for the current processing type in experiment 2 for partition dependence in Section 6.4.

Item	Construct	Loading	Coding
iPhone			
Durch Erfahrungen mit vergleichbaren Situationen in der Vergangenheit fällt es mir leicht eine Preiseinschätzung in dieser Situation abzugeben.	Heuristic processing	positive	Likert-7
Bezüglich des wahrscheinlichen Preises des iPhone 7S in Deutschland bei dessen Markteinführung bin ich bereit, mich auf Experten zu verlassen.	Heuristic processing	positive	Likert-7

TABLE A.5: Items for the context specific questionnaire for the current processing type in experiment 2 for partition dependence in Section 6.4.

Item	**Construct**	**Loading**	**Coding**
Ich konnte eine Entscheidung darüber treffen, wie ich den wahrscheinlichen Preis des iPhone 7S in Deutschland bei dessen Markteinführung einschätze, ohne eine große Menge zusätzlicher Informationen zu suchen, indem ich mein vorhandenes Wissen genutzt habe.	Heuristic processing	positive	Likert-7
Um vollständig bezüglich des wahrscheinlichen Preises des iPhone 7S in Deutschland bei dessen Markteinführung informiert zu sein, habe ich das Gefühl, dass ich besser dran bin, je mehr Anhaltspunkte ich bekommen kann.	Systematic processing	positive	Likert-7
Ich bemühe mich stark, sorgfältig die relevanten Informationen bezüglich des wahrscheinlichen Preises von neuen iPhones bei deren Markteinführung herauszufinden.	Systematic processing	positive	Likert-7
Wenn das Thema des wahrscheinlichen Preises des iPhone 7S in Deutschland bei dessen Markteinführung aufkommt, versuche ich immer mehr darüber zu erfahren.	Systematic processing	positive	Likert-7
Wenn ich auf Informationen zum Thema des wahrscheinlichen Preises des iPhone 7S in Deutschland bei dessen Markteinführung stoße, ist es wahrscheinlich, dass ich innehalte und sorgfältig darüber nachdenke.	Systematic processing	positive	Likert-7
Coalitions			
Durch Erfahrungen mit vergleichbaren Situationen in der Vergangenheit fällt es mir leicht eine Entscheidung in dieser Situation zu treffen.	Heuristic processing	positive	Likert-7
Bezüglich der wahrscheinlichen Koalitionsbildung nach der Bundestagswahl 2017 bin ich bereit, mich auf Experten zu verlassen.	Heuristic processing	positive	Likert-7

TABLE A.5: Items for the context specific questionnaire for the current processing type in experiment 2 for partition dependence in Section 6.4.

Item	**Construct**	**Loading**	**Coding**
Ich konnte eine Entscheidung darüber treffen, wie ich die wahrscheinliche Koalitionsbildung nach der Bundestagswahl 2017 einschätze, ohne eine große Menge zusätzlicher Informationen zu suchen, indem ich mein vorhandenes Wissen genutzt habe.	Heuristic processing	positive	Likert-7
Um vollständig bezüglich der wahrscheinlichen Koalitionsbildung nach der Bundestagswahl 2017 informiert zu sein, habe ich das Gefühl, dass ich besser dran bin, je mehr Anhaltspunkte ich bekommen kann.	Systematic processing	positive	Likert-7
Ich habe mich stark bemüht, sorgfältig die wissenschaftlichen Informationen bezüglich der wahrscheinlichen Koalitionsbildung nach der Bundestagswahl 2017 herauszufinden.	Systematic processing	positive	Likert-7
Wenn das Thema der wahrscheinlichen Koalitionsbildung nach der Bundestagswahl 2017 aufkommt, versuche ich immer mehr darüber zu erfahren.	Systematic processing	positive	Likert-7
Wenn ich auf Informationen zum Thema der wahrscheinlichen Koalitionsbildung nach der Bundestagswahl 2017 stoße, ist es wahrscheinlich, dass ich innehalte und sorgfältig darüber nachdenke.	Systematic processing	positive	Likert-7

TABLE A.6: Items for the expertise questionnaire in experiment 2 for partition dependence in Section 6.4.

Item	**Construct**	**Loading**	**Coding**
Könntest du jemandem viel über dieses Thema erzählen?	Expertise	positive	Likert-7
Würdest du darüber einen Zeitungsartikel lesen?	Expertise	positive	Likert-7
Denkst du, dass Freude dich zu diesem Thema befragen würden?	Expertise	positive	Likert-7
Hat dich dieses Thema schon mal betroffen?	Expertise	positive	Likert-7

Bibliography

Abadie, Fabienne, Michael Friedewald, and Matthias Weber (2010). "Adaptive foresight in the creative content industries: anticipating value chain transformations and need for policy action". In: *Science and Public Policy* 37.1, p. 12. DOI: 10.3152/030234210X484793.

Abramowicz, Michael (2004). "Information Markets, Administrative Decisionmaking, and Predictive Cost-Benefit Analysis". In: *The University of Chicago Law Review* 71.3, pp. 933–1020. ISSN: 00419494. URL: http://www.jstor.org/stable/1600601.

Admati, Anat R and Paul Pfleiderer (1988). "A theory of intraday patterns: Volume and price variability". In: *The Review of Financial Studies* 1.1, pp. 3–40. ISSN: 0893-9454.

Allen, Franklin and Douglas Gale (1992). "Stock-price manipulation". In: *Review of financial studies* 5.3, pp. 503–529. ISSN: 0893-9454.

Almendra, Vinicius and Daniel Schwabe (2009). "Fraud Detection by Human Agents: A Pilot Study BT - E-Commerce and Web Technologies". In: *Proceedings of the 10th International Conference, EC-Web 2009, Linz, Austria, September 1-4, 2009.* Ed. by Tommaso Di Noia and Francesco Buccafurri. Springer Berlin Heidelberg, pp. 300–311. ISBN: 978-3-642-03964-5. DOI: 10.1007/978-3-642-03964-5_28.

Ames, Morgan and Mor Naaman (2007). "Why We Tag: Motivations for Annotation in Mobile and Online Media". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. New York, NY, USA: ACM, pp. 971–980. ISBN: 978-1-59593-593-9. DOI: 10.1145/1240624.1240772.

Ammon, Ursula (2009). "Delphi-Befragung". In: *Handbuch Methoden der Organisationsforschung: Quantitative und Qualitative Methoden*. Ed. by Stefan Kühl, Petra Strodtholz, and Andreas Taffertshofer. Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 458–476. ISBN: 978-3-531-91570-8. DOI: 10.1007/978-3-531-91570-8_22.

Anderson, Robert Whitney and Cathy Ross (2014). *Collaborative Fraud Determination And Prevention*. URL: https://www.google.com/patents/US20140108251.

Armstrong, J Scott (1985). "Judgmental Methods". In: *Long-Range Forecasting: From Crystal Ball to Computer*. 2nd ed. John Wiley & Sons Inc. Chap. 6, pp. 79–149.

Armstrong, J Scott and Kesten C Green (2018). "Forecasting Methods and Principles: Evidence-Based Checklists". In: *Journal of Global Scholars of Marketing Science* 28.2, pp. 103–159. DOI: 10.1080/21639159.2018.1441735.

Arrow, Kenneth J et al. (2008). "The Promise of Prediction Markets". In: *Science* 320.5878, pp. 877–878. DOI: 10.1126/science.1157679.

Atanasov, Pavel et al. (2016). "Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls". In: *Management Science* 63.3, pp. 691–706. ISSN: 0025-1909. DOI: 10.1287/mnsc.2015.2374.

Aulerich, Nicole M, Scott H Irwin, and Philip Garcia (2013). "Returns to individual traders in agricultural futures markets: skill or luck?" In: *Applied Economics* 45.25, pp. 3650–3666. ISSN: 0003-6846. DOI: 10.1080/00036846.2012.727979.

Baba, Yukino et al. (2013). "Leveraging Crowdsourcing to Detect Improper Tasks in Crowdsourcing Marketplaces." In: *Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference*, pp. 1487–1492.

Bader, Max (2013). "Crowdsourcing election monitoring in the 2011–2012 Russian elections". In: *East European Politics* 29.4, pp. 521–535. ISSN: 2159-9165. DOI: 10.1080/21599165.2013.818979.

Bagnoli, Mark and Barton L Lipman (1996). "Stock price manipulation through takeover bids". In: *The RAND Journal of Economics* 27.1, pp. 124–147. ISSN: 0741-6261. DOI: 10.2307/2555795.

Ban, Amir (2018). "Strategy-Proof Incentives for Predictions". URL: http://arxiv.org/abs/1805.04867.

Barclay, Michael J and Jerold B Warner (1993). "Stealth trading and volatility: Which trades move prices?" In: *Journal of Financial Economics* 34.3, pp. 281–305. ISSN: 0304-405X.

Bardolet, David, Craig R Fox, and Daniel Lovallo (2011). "Naïve diversification and partition dependence in capital allocation decisions: field and experimental evidence". In: *Strategic Management Journal* 32, pp. 1465–1483. DOI: 10.1002/smj.966.

Baron, Reuben M. and David a. Kenny (1986). "The Moderator-Mediator Variable Distinction in Social The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations". In: *Journal of Personality and Social Psychology* 51.6, pp. 1173–1182. ISSN: 0022-3514. DOI: 10.1037/0022-3514.51.6.1173. arXiv: 4.

Benabou, Roland and Guy Laroque (1992). "Using privileged information to manipulate markets: Insiders, gurus, and credibility". In: *The Quarterly Journal of Economics* 107.3, pp. 921–958. ISSN: 0033-5533. DOI: 10.2307/2118369.

Benson, Buster (2016). *Better Humans: Cognitive bias cheat sheet*. URL: https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18 (visited on 05/28/2018).

Berg, Joyce E, Robert Forsythe, and Thomas A Rietz (1997). "What makes markets predict well? Evidence from the Iowa Electronic Markets". In: *Understanding Strategic Interaction*. Ed. by Wulf Albers et al. Springer, pp. 444–463.

Berg, Joyce E, Forrest D Nelson, and Thomas A Rietz (2008). "Prediction Market Accuracy in the Long Run". In: *The International Journal of Forecasting* 24.January, pp. 285–300. DOI: 10.1016/j.ijforecast.2008.03.007.

Berg, Joyce E and Thomas A Rietz (2018). "Longshots , Overconfidence and Efficiency on the Iowa Electronic Market". In: *International Journal of Forecasting*, to appear. URL: https://www.biz.uiowa.edu/faculty/trietz/papers/Longshots.pdf.

Berlemann, Michael and Carsten Schmidt (2001). "Predictive accuracy of political stock markets: Empirical evidence from a European perspective". In: *Dresden Discussion Paper Series in Economics 05/01,* Dresden: Technische Universität Dresden, Faculty of Business and Economics, Department of Economics.

Best, Roger J (1974). "An Experiment in Delphi Estimation in Marketing Decision Making". In: *Journal of Marketing Research* 11.4, pp. 448–452. ISSN: 00222437. DOI: 10.2307/3151295.

Bhattacherje, Anol (2001). "Understanding Information Systems Continuance: An Expectation-Confirmation Model". In: *MIS Quarterly* 25.3, pp. 351–370.

Blanc, Sebastian M. and Thomas Setzer (2016). "When to choose the simple average in forecast combination". In: *Journal of Business Research* 69.10, pp. 3951–3962. ISSN: 01482963. DOI: 10.1016/j.jbusres.2016.05.013.

Blume, Michael (2012). "Behavior identification in markets using visualization and network analysis". PhD thesis. Karlsruhe Institute of Technology. URL: http://digbib.ubka.uni-karlsruhe.de/volltexte/1000026214.

Blume, Michael, Stefan Luckner, and Christof Weinhardt (2010). "Fraud detection in play-money prediction markets". In: *Information Systems and E-Business Management* 8.4, pp. 395–413. ISSN: 1617-9846.

Bohm, Peter and Joakim Sonnegard (1999). "Political stock markets and unreliable polls". In: *The Scandinavian Journal of Economics* 101.2, pp. 205–222. ISSN: 1467-9442.

Bolger, Fergus and George Wright (2011). "Improving the Delphi process: Lessons from social psychological research". In: *Technological Forecasting and Social Change* 78.9, pp. 1500–1513. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2011.07.007.

Bothos, Efthimios, Dimitris Apostolou, and Gregoris Mentzas (2009). "IDEM: A Prediction Market for Idea Management". In: *WEB2008: Designing E-Business Systems. Markets, Services, and Networks*. Ed. by Christof Weinhardt, Stefan Luckner, and Jochen Stößer. Berlin, Heidelberg: Springer Berlin Heidelberg. Chap. 1, pp. 1–13. ISBN: 978-3-642-01256-3.

Boulkedid, Rym et al. (2011). "Using and Reporting the Delphi Method for Selecting Healthcare Quality Indicators: A Systematic Review". In: *PLoS ONE* 6.6, e20476. DOI: 10.1371/journal.pone.0020476.

Brakel, Linda A W, Howard Shevrin, and Karen K Villa (2002). "The priority of primary process categorizing: Experimental evidence supporting a psychoanalytic developmental hypothesis". In: *Journal of the American Psychoanalytic Association* 50.2, pp. 483–505. ISSN: 0003-0651.

Brug, Wouter van der, Meindert Fennema, and Jean Tillie (2000). "Anti-immigrant Parties in Europe: Ideological or Protest Vote?" In: *European Journal of Political Research* 37.1, pp. 77–102. ISSN: 1475-6765. DOI: 10.1111/1475-6765.00505.

Brüggelambert, Gregor (1999). "Institutionen als Informationsträger: Erfahrungen mit Wahlbörsen". PhD thesis. Marburg: University of Essen, Germany. ISBN: 38951-82346.

Bruneel, Christophe et al. (2018). "Movie Analytics and the Future of Film Finance . Are Oscars and Box Office Revenue Predictable ?" In: *Handbook of State Aid for Film*. Ed. by Paul Clemens Murschetz, Roland Teichmann, and Matthias Karmasin. Springer International Publishing AG, pp. 551–578. ISBN: 9783319717166. DOI: 10.1007/978-3-319-71716-6.

Brünken, Roland, Jan L Plass, and Detlev Leutner (2003). "Direct Measurement of Cognitive Load in Multimedia Learning". In: *Educational Psychologist* 38.1, pp. 53–61. DOI: 10.1207/S15326985EP3801_7.

Buckley, Patrick (2016). "Harnessing the wisdom of crowds: Decision spaces for prediction markets". In: *Business Horizons* 59.1, pp. 85–94. ISSN: 0007-6813. DOI: 10.1016/j.bushor.2015.09.003.

Buckley, Patrick and Elaine Doyle (2015). "Using web-based collaborative forecasting to enhance information literacy and disciplinary knowledge". In: *Interactive Learning Environments* 24.7, pp. 1574–1589. ISSN: 17445191. DOI: 10.1080/10494820.2015.1041399.

— (2017). "Individualising gamification: An investigation of the impact of learning styles and personality traits on the efficacy of gamification using a prediction market". In: *Computers and Education* 106, pp. 43–55. ISSN: 03601315. DOI: 10.1016/j.compedu.2016.11.009.

Budescu, David V and Eva Chen (2017). "Identifying Expertise to Extract the Wisdom of Crowds". In: *Management Science* 61.2, pp. 267–280. DOI: 10.1287/mnsc.2014.1909.

Camerer, Colin F (1998). "Can Asset Markets Be Manipulated? A Field Experiment with Racetrack Betting". In: *Journal of Political Economy* 106.3, pp. 457–482. ISSN: 0022-3808. DOI: 10.1086/250018.

Campbell, Sean D, Steven A Sharpe, and Others (2009). "Anchoring bias in consensus forecasts and its effect on market prices". In: *Journal of Financial and Quantitative Analysis* 44.2, p. 369.

Candy, Linda (1998). "Creative knowledge work and interaction design". PhD thesis. Loughborough University. URL: https://dspace.lboro.ac.uk/2134/6992.

Candy, Linda and Ernest A Edmonds (1995). "Creativity in Knowledge Work: A Process Model and Requirements for Support". In: *Proceedings OZCHI 1995*, pp. 242–248.

Carvalho, Arthur (2017). "On a participation structure that ensures representative prices in prediction markets". In: *Decision Support Systems* 104, pp. 13–25. ISSN: 01679236. DOI: 10.1016/j.dss.2017.09.008.

Casas, Augustin, Yarine Fawaz, and Andre Trindade (2016). "Surprise Me If You Can: The Influence of Newspaper Endorsements in U.S. Presidential Elections". In: *Economic Inquiry* 54.3, pp. 1484–1498. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2777778.

Chakraborty, Mithun and Sanmay Das (2016). "Trading on a rigged game: outcome manipulation in prediction markets". In: *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 158–164. URL: http://dl.acm.org/citation.cfm?id=3060644.

Chakravarty, Sugato (2001). "Stealth-trading: Which traders' trades move stock prices?" In: *Journal of Financial Economics* 61.2, pp. 289–307. ISSN: 0304-405X.

Chen, Ouhao et al. (2017). "Extending Cognitive Load Theory to Incorporate Working Memory Resource Depletion: Evidence from the Spacing Effect". In: *Educational Psychology Review* 30.2, pp. 483–501. ISSN: 1573-336X. DOI: 10.1007/s10648-017-9426-2.

Chen, Weiyun, Xin Li, and D D Zeng (2015). "Simple is beautiful: Toward light prediction markets". In: *Intelligent Systems, IEEE* 30.3, pp. 76–80. ISSN: 1541-1672. DOI: 10.1109/MIS.2015.54. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7111878.

Chen, Yiling and Ian A Kash (2011). "Information Elicitation for Decision Making". In: *The 10th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 175–182. URL: http://dl.acm.org/citation.cfm?id=2030470.2030496.

Chen, Yiling and David M Pennock (2010). "Designing Markets for Prediction". In: *AI Magazine* 31.4, pp. 42–52.

Chen, Yiling et al. (2007). "Bluffing and Strategic Reticence in Prediction Markets". In: *Internet and Network Economics*. Ed. by Xiaotie Deng and Fan Chung Graham. Berlin, Heidelberg: Springer Berlin Heidelberg. Chap. 10, pp. 70–81. ISBN: 978-3-540-77105-0. DOI: 10.1007/978-3-540-77105-0_10.

Chen, Yiling et al. (2015). "Market Manipulation with Outside Incentives". In: *Autonomous Agents and Multi-Agent Systems* 29.2, pp. 230–265. ISSN: 1387-2532. DOI: 10.1007/s10458-014-9249-1.

Cipriano, Michael C and Thomas S Gruca (2014). "The Power of Priors: How Confirmation Bias impacts Market Prices". In: *Journal of Prediction Markets* 8.3, pp. 34–56. ISSN: 17506751.

Clayton, Mark J (1997). "Delphi: a technique to harness expert opinion for critical decision-making tasks in education". In: *Educational Psychology* 17.4, pp. 373–386. ISSN: 0144-3410. DOI: 10.1080/0144341970170401.

Cohen, Jacob (1988). *Statistiacal Power Analysis for the Behavioural Sciences*. 2nd ed. Lawrence Erlbaum Associates, p. 579. ISBN: 0805802835.

Cowgill, Bo, Justin Wolfers, and Eric Zitzewitz (2009). "Using Prediction Markets to Track Information Flows: Evidence from Google". In: *1st International Conference on Auctions, Market Mechanisms and Their Applications 2009*. Ed. by Sanmay Das et al. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 3. ISBN: 978-3-642-03821-1. DOI: 10.1007/978-3-642-03821-1_2.

Cressey, Donald R (1953). *Other people's money; a study of the social psychology of embezzlement.* New York, NY, US: Free Press, p. 191. ISBN: 978-0534001421. URL: http://psycnet.apa.org/record/1954-06293-000.

Cuhls, Kerstin (2003). "From forecasting to foresight processes—new participative foresight activities in Germany". In: *Journal of Forecasting* 22.2-3, pp. 93–111. ISSN: 1099-131X. DOI: 10.1002/for.848.

Curtis, Andrew and Rachel Wood (2004). "Optimal elicitation of probabilistic information from experts". In: *Geological Society, London, Special Publications* 239.1, 127 LP –145. URL: http://sp.lyellcollection.org/content/239/1/127.abstract.

Dahan, Ely et al. (2011). "Securities Trading of Concepts (STOC)". In: *Journal of Marketing Research* 48.3, pp. 497–517. ISSN: 0022-2437. DOI: 10.1509/jmkr.48.3.497.

Dalkey, Norman Crolee, Bernice B Brown, and Samuel Cochran (1969). *The Delphi method: An experimental study of group opinion*. Vol. 3. Rand Corporation Santa Monica, CA. URL: https://www.rand.org/content/dam/rand/pubs/research_memoranda/2005/RM5888.pdf.

Darkow, Inga-Lena and Heiko A von der Gracht (2013). "Scenarios for the future of the European process industry - the case of the chemical industry". In: *European Journal of Futures Research* 1.1, pp. 1–12. ISSN: 2195-2248. DOI: 10.1007/s40309-013-0010-9.

Deck, Cary, Shengle Lin, and David Porter (2013). "Affecting policy by manipulating prediction markets: Experimental evidence". In: *Journal of Economic Behavior & Organization* 85, pp. 48–62. ISSN: 0167-2681.

Diemer, Sebastian and Joaquin Poblete (2010). "Real-Money Vs. Play-Money Forecasting Accuracy in Online Prediction Markets - Empirical Insights from Ipredict". In: *Journal of Prediction Markets* 4.3, pp. 21–58.

Dimitrov, Stanko and Rahul Sami (2010). "Composition of Markets with Conflicting Incentives". In: *Proceedings of the 11th ACM Conference on Electronic Commerce*. ACM, pp. 53–62. DOI: 10.1145/1807342.1807350.

Dohmen, Thomas et al. (2011). "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences". In: *Journal of the European Economic Association* 9.3, pp. 522–550. ISSN: 1542-4766. DOI: 10.1111/j.1542-4774.2011.01015.x.

Dreber, Anna et al. (2015). "Using prediction markets to estimate the reproducibility of scientific research". In: *Proceedings of the National Academy of Sciences* 112.50, pp. 15343–15347. ISSN: 0027-8424. DOI: 10.1073/pnas.1516179112.

Dresch, Aline, Daniel Pacheco Lacerda, and José Antônio Valle Antunes (2015). "Class of Problems and Artifacts". In: *Design Science Research: A Method for Science and Technology Advancement*. Ed. by Aline Dresch, Daniel Pacheco Lacerda, and José Antônio Valle Antunes Jr. Cham: Springer International Publishing. Chap. 5, pp. 103–116. ISBN: 978-3-319-07374-3. DOI: 10.1007/978-3-319-07374-3_5.

Durand, Rodolphe (2003). "Predicting a firm's forecasting ability: the roles of organizational illusion of control and organizational attention". In: *Strategic Management Journal* 24.9, pp. 821–838. ISSN: 1097-0266. DOI: 10.1002/smj.339.

Easley, David and Maureen O'hara (1987). "Price, trade size, and information in securities markets". In: *Journal of Financial economics* 19.1, pp. 69–90. ISSN: 0304-405X.

Eickhoff, Carsten (2018). "Cognitive Biases in Crowdsourcing". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. New York, NY, USA: ACM, pp. 162–170. ISBN: 978-1-4503-5581-0. DOI: 10.1145/3159652.3159654.

Epstein, Seymour et al. (1996). "Individual Differences in Intuitive–Experiential and Analytical–Rational Thinking Styles". In: *Journal of Personality and Social Psychology* 72.2, pp. 390–405. DOI: 10.1037/0022-3514.71.2.390.

Erat, Sanjiv and Uri Gneezy (2016). "Incentives for creativity". In: *Experimental Economics* 19.2, pp. 269–280. ISSN: 1573-6938. DOI: 10.1007/s10683-015-9440-5.

Erikson, Robert S and Christopher Wlezien (2012). "Markets vs. polls as election predictors: An historical assessment". In: *Electoral Studies* 31.3, pp. 532–539. ISSN: 0261-3794. DOI: 10.1016/j.electstud.2012.04.008.

Evans, Jonathan St. B. T. (2012). "Dual process theories of deductive reasoning: facts and fallacies". In: *The Oxford Handbook of Thinking and Reasoning*. Ed. by Keith J. Holyoak and Robert G. Morrison. OUP USA, 2013. Chap. 8, pp. 155–133. ISBN: 0199313792, 9780199313792.

Fama, Eugene F (1970). "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *The Journal of Finance* 25.2, pp. 383–417. ISSN: 00221082, 15406261. DOI: 10.2307/2325486.

Feess, Eberhard, Helge Müller, and Christoph Schumacher (2014). "The favorite-longshot bias and the impact of experience". In: *Business Research* 7.2, pp. 217–234. ISSN: 2198-2627. DOI: 10.1007/s40685-014-0013-9.

Fischhoff, B, P Slovic, and S Lichtenstein (1978). "Fault trees: Sensitivity of estimated failure probabilities to problem representation." In: *Journal of Experimental Psychology: Human Perception and Performance* 4.2, pp. 330–344. DOI: 10.1037/0096-1523.4.2.330.

Forsythe, Robert, Thomas A Rietz, and Thomas W Ross (1999). "Wishes, expectations and actions: a survey on price formation in election stock markets". In: *Journal of Economic Behavior & Organization* 39.1, pp. 83–110. ISSN: 0167-2681.

Fox, Craig R, David Bardolet, and Daniel Lieb (2005). "Partition dependence in decision analysis, resource allocation, and consumer choice". In: *Experimental business research*. Ed. by R Zwick and A Rapoport. Dordrecht: The Netherlands: Kluwer, pp. 338–360.

Fox, Craig R and Robert T. Clemen (2005). "Subjective Probability Assessment in Decision Analysis: Partition Dependence and Bias Toward the Ignorance Prior". In: *Management Science* 51.9, pp. 1417–1432. ISSN: 0025-1909. DOI: 10.1287/mnsc.1050.0409.

Franke, Markus, Andreas Geyer-Schulz, and Bettina Hoser (2005). "Analyzing Trading Behavior in Transaction Data of Electronic Election Markets". In: *Data Analysis and Decision Support*. Ed. by Daniel Baier, Reinhold Decker, and Lars Schidt-Thieme. Berlin, Heidelberg: Springer, pp. 222–230.

— (2006). "On the Analysis of Asymmetric Directed Communication Structures in Electronic Election Markets". In: *Agent-Based Computational Modelling: Applications in Demography, Social, Economic and Environmental Sciences*. Ed. by Francesco C Billari et al. Heidelberg: Physica-Verlag HD, pp. 37–59. ISBN: 978-3-7908-1721-8. DOI: 10.1007/3-7908-1721-X_3.

Franke, Markus, Bettina Hoser, and Jan Schröder (2008). "On the Analysis of Irregular Stock Market Trading Behavior". In: *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*. Ed. by Christine Preisach et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 355–362. ISBN: 978-3-540-78246-9. DOI: 10.1007/978-3-540-78246-9_42.

Frey, Bruno S and Reto Jegen (2001). "Motivation crowding theory". In: *Journal of economic surveys* 15.5, pp. 589–611. ISSN: 1467-6419.

Fry, John and Andrew Brint (2017). "Bubbles, Blind-Spots and Brexit". In: *Risks* 5.37, pp. 1–15. DOI: 10.3390/risks5030037.

Galton, Francis (1907). "Vox populi (The wisdom of crowds)". In: *Nature* 75.7, pp. 450–451.

Gangur, Mikuláš (2016). "Motivation System on Prediction Market". In: *Proceedings of the 8th International Conference on Computational Collective Intelligence, ICCCI 2016, Halkidiki, Greece, September 28-30, 2016, Part II*. Ed. by Ngoc Thanh Nguyen et al. Cham: Springer International Publishing. Chap. 10, pp. 354–363. ISBN: 978-3-319-45246-3. DOI: 10.1007/978-3-319-45246-3_34.

Garcia-Molina, H et al. (2016). "Challenges in Data Crowdsourcing". In: *IEEE Transactions on Knowledge and Data Engineering* 28.4, pp. 901–911. ISSN: 1041-4347 VO - 28. DOI: 10.1109/TKDE.2016.2518669.

Garcin, Florent and Boi Faltings (2014). "Swissnoise: Online Polls with Game-Theoretic Incentives". In: *Proceedings of the Twenty-Sixth Annual Conference on Innovative Applications of Artificial Intelligence*, pp. 2972–2977.

Gary, Jay E. and Heiko A von der Gracht (2015). "The future of foresight professionals: Results from a global Delphi study". In: *Futures* 71, pp. 132–145. ISSN: 00163287. DOI: 10.1016/j.futures.2015.03.005.

Gefen, David and Detmar W Straub (2004). "Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services". In: *Omega* 32.6, pp. 407–424. ISSN: 0305-0483. DOI: 10.1016/j.omega.2004.01.006.

George Mason University (2015). *SciCast Annual Report (2015)*. Tech. rep. George Mason University. URL: https://mason.gmu.edu/~rhanson/SciCast2015.pdf.

Ghosh, Arpita, Satyen Kale, and Preston McAfee (2011). "Who Moderates the Moderators?: Crowdsourcing Abuse Detection in User-generated Content". In: *Proceedings of the 12th ACM Conference on Electronic Commerce*. EC '11. New York, NY, USA: ACM, pp. 167–176. ISBN: 978-1-4503-0261-6. DOI: 10.1145/1993574.1993599.

Gimpel, Henner (2007). "Loss aversion and reference-dependent preferences in multi-attribute negotiations". In: *Group Decision and Negotiation* 16.4, pp. 303–319. ISSN: 09262644. DOI: 10.1007/s10726-006-9051-9.

Gnatzy, Tobias et al. (2011). "Validating an innovative real-time Delphi approach - A methodological comparison between real-time and conventional Delphi studies". In: *Technological Forecasting and Social Change* 78.9, pp. 1681–1694. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2011.04.006.

Gómez-Chacón, Inés Mª et al. (2014). "The dual processes hypothesis in mathematics performance: Beliefs, cognitive reflection, working memory and reasoning". In: *Learning and Individual Differences* 29, pp. 67–73. ISSN: 10416080. DOI: 10.1016/j.lindif.2013.10.001.

Goodman, Claire (2017). "Conversation or consensus: Using the Delphi technique to set priorities for ageing research and practice". In: *Age and Ageing* 46.1, pp. 6–7. ISSN: 14682834. DOI: 10.1093/ageing/afw183.

Gordon, Theodore J (2007). "Energy forecasts using a "Roundless" approach to running a Delphi study". In: *Foresight* 9.2, pp. 27–35. ISSN: 1463-6689. DOI: 10.1108/14636680710737731.

— (2009). *The real-time Delphi method*. Tech. rep. The Millenium Project, pp. 1–21. URL: http://www.econ.uba.ar/unai/archivos/actividades/30_de_octubre/Real_Time_Delphi_Ted_Gordon.pdf.

Gordon, Theodore J and Olaf Helmer (1964). *Report on a long-range forecasting study*. Tech. rep. THE RAND CORPERATION, p. 65. DOI: 10.1126/science.1088667. URL: http://www.rand.org/content/dam/rand/pubs/papers/2005/P2982.pdf.

Gordon, Theodore J and Adam Pease (2006). "RT Delphi: An efficient, "round-less" almost real time Delphi method". In: *Technological Forecasting and Social Change* 73.4, pp. 321–333. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2005.09.005.

Gordon, Theodore J, Yair Sharan, and Elizabeth Florescu (2015). "Prospects for Lone Wolf and SIMAD terrorism". In: *Technological Forecasting and Social Change* 95, pp. 234–251. ISSN: 00401625. DOI: 10.1016/j.techfore.2015.01.013.

Graefe, Andreas (2011). "Prediction market accuracy for business forecasting". In: *Prediction Markets: Theory and Applications*. Ed. by Leighton Vaughan Williams. Routledge. Chap. 7, pp. 87–95. ISBN: 9781136715693.

— (2014). "Accuracy of Vote Expectation Surveys in Forecasting Elections". In: *Public Opinion Quarterly* 78.S1, pp. 204–232. ISSN: 0033-362X. DOI: 10.1093/poq/nfu008.

— (2015). "German Election Forecasting: Comparing and Combining Methods for 2013". In: *German Politics* 24.2, pp. 195–204. DOI: 10.1080/09644008.2015.1024240.

— (2017). "Prediction Market Performance in the 2016 U.S. Presidential Election". In: *Foresight: The International Journal of Applied Forecasting* 1.45, pp. 38–42. URL: http://econpapers.repec.org/RePEc:for:ijafaa:y:2017:i:45:p:38-42.

Graefe, Andreas and J Scott Armstrong (2011). "Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task". In: *International Journal of Forecasting* 27.1, pp. 183–195. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2010.05.004.

Graefe, Andreas, Stefan Luckner, and Christof Weinhardt (2010). "Prediction markets for foresight". In: *Futures* 42.4, pp. 394–404. ISSN: 00163287. DOI: 10.1016/j.futures.2009.11.024.

Graefe, Andreas et al. (2015). "Limitations of Ensemble Bayesian Model Averaging for forecasting social science problems". In: *International Journal of Forecasting* 31.3, pp. 943–951. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2014.12.001.

Green, Kesten C, J Scott Armstrong, and Andreas Graefe (2007). "Methods to elicit forecasts from groups: Delphi and prediction markets compared". In: *Foresight Int. J. Appl. Forecast* 8.17-20.

Greer, Jennifer D and Mark E LaPointe (2004). "Cyber-campaigning grows up: A comparative content analysis of websites for US Senate and gubernatorial races, 1998-2000". In: *Electronic Democracy: Mobilisation, organisation and participation via new ICTs*. Ed. by Rachel K Gibson, Andrea Römmele, and Stephen J Ward. Routledge. Chap. 6, pp. 116–132.

Gregor, Shirley (2006). "The Nature of Theory in Information Systems". In: *MIS Q.* 30.3, pp. 611–642. ISSN: 0276-7783. URL: http://dl.acm.org/citation.cfm?id=2017296.2017300.

Gruca, Thomas S and Joyce E Berg (2007). "Public Information Bias and Prediction Market Accuracy". In: *The Journal of Prediction Markets* 1.3, pp. 219–231. DOI: 10.5750/jpm.v1i3.430.

Gruca, Thomas S, Joyce E Berg, and Michael C Cipriano (2008). "Incentive and accuracy issues in movie prediction markets". In: *Journal of Prediction Markets* 2.1, pp. 29–43.

Hall, Caitlin (2010). "Prediction Markets: Issues and Applications". In: *Journal of Prediction Markets* 4.1, pp. 27–58. DOI: 10.5750/jpm.v4i1.472.

Hansen, Jan, Carsten Schmidt, and Martin Strobel (2004). "Manipulation in political stock markets: preconditions and evidence". In: *Applied Economics Letters* 11.7, pp. 459–463. ISSN: 1350-4851.

Hanson, Robin (2002). "Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation". In: *Journal of Prediction Markets* 1.1, pp. 3–15. URL: http://www.ubplj.org/index.php/jpm/article/view/417.

— (2006a). "Designing real terrorism futures". In: *Public Choice* 128.1, pp. 257–274. ISSN: 1573-7101. DOI: 10.1007/s11127-006-9053-9.

— (2006b). "Foul play in information markets". In: *Information markets: A new Way of Making Decisions*. Ed. by Robert W Hahn and Paul C Tetlock. AEI Press. Chap. 6, pp. 126–141.

Hanson, Robin and Ryan Oprea (2009). "A Manipulator Can Aid Prediction Market Accuracy". In: *Economica* 76.302, pp. 304–314. ISSN: 1468-0335. DOI: 10.1111/j.1468-0335.2008.00734.x.

Hanson, Robin, Ryan Oprea, and David Porter (2006). "Information aggregation and manipulation in an experimental market". In: *Journal of Economic Behavior & Organization* 60.4, pp. 449–459. ISSN: 0167-2681.

Hardford, Tim (2007). *Undercover Economist: Tote that vote*. URL: https://www.ft.com/content/242d5378-22c4-11dc-ac53-000b5df10621 (visited on 04/06/2017).

Harris, Christopher (2012). "Detecting deceptive opinion spam using human computation". In: *Workshops at AAAI 2012 on Artificial Intelligence*, pp. 87–93.

Harris, Larry (2003). *Trading and exchanges: Market microstructure for practitioners*. Oxford University Press, USA. ISBN: 0195144708.

Harvey, Nigel (2007). "Use of heuristics: Insights from forecasting research". In: *Thinking & Reasoning* 13.1, pp. 5–24. ISSN: 1354-6783.

Hasson, Felicity and Sinead Keeney (2011). "Enhancing rigour in the Delphi technique research". In: *Technological Forecasting and Social Change* 78.9, pp. 1695–1704. ISSN: 00401625. DOI: 10.1016/j.techfore.2011.04.005.

Hayek, F A (1945). "The use of knowledge in society". In: *American Economic Review* 35.4, pp. 519–530.

Heckman, James (1990). "Varieties of Selection Bias". In: *The American Economic Review* 80.2, pp. 313–318.

Hevner, Alan R (2007). "A three cycle view of design science research". In: *Scandinavian journal of information systems* 19.2, pp. 1–6.

Hevner, Alan R et al. (2004). "Design Science in Information Systems Research". In: *MIS Quarterly* 28.1, pp. 75–105. ISSN: 02767783. URL: http://www.jstor.org/stable/25148625.

Hill, Kim Quaile and Jib Fowles (1975). "The methodological worth of the Delphi forecasting technique". In: *Technological Forecasting and Social Change* 7.2, pp. 179–192. ISSN: 0040-1625. DOI: 10.1016/0040-1625(75)90057-8.

Hirth, Matthias, Tobias Hoßfeld, and Phuoc Tran-Gia (2011). "Cost-Optimal Validation Mechanisms and Cheat-Detection for Crowdsourcing Platforms". In: *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on*. Seoul, South Korea, pp. 316–321. DOI: 10.1109/IMIS.2011.91.

Hirth, Matthias, Tobias Hoßfeld, and Phuoc Tran-Gia (2013). "Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms". In: *Mathematical and Computer Modelling* 57.11, pp. 2918–2932. ISSN: 0895-7177. DOI: 10.1016/j.mcm.2012.01.006. URL: http://www.sciencedirect.com/science/article/pii/S0895717712000076.

Hsu, Chin-Lung and Hsi-Peng Lu (2004). "Why do people play on-line games? An extended TAM with social influences and flow experience". In: *Information & Management* 41.7, pp. 853–868. ISSN: 0378-7206. DOI: 10.1016/j.im.2003.08.014. URL: http://www.sciencedirect.com/science/article/pii/S0378720603001319.

Huang, Eric H and Yoav Shoham (2014). "Price Manipulation in Prediction Markets: Analysis and Mitigation". In: *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*. AAMAS '14. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, pp. 213–220. ISBN: 978-1-4503-2738-1. URL: http://dl.acm.org/citation.cfm?id=2615731.2615768.

Huang, Eric Hsin-Chun (2016). "Maintaining the ownership of information in the digital age". PhD thesis. Stanford University. URL: https://searchworks.stanford.edu/view/11602319.

Ishikawa, Akira et al. (1993). "The max-min Delphi method and fuzzy Delphi method via fuzzy integration". In: *Fuzzy Sets and Systems* 55.3, pp. 241–253. ISSN: 0165-0114. DOI: 10.1016/0165-0114(93)90251-C.

Jian, Lian and Rahul Sami (2012). "Aggregation and Manipulation in Prediction Markets: Effects of Trading Mechanism and Information Distribution". In: *Management Science* 58.1, pp. 123–140.

Jones, Robert C (2014). "Making Better (Investment) Decisions." In: *Journal of Portfolio Management* 40.2, pp. 128–143. ISSN: 00954918.

Judgmental Forecasting (2009). In: *Farlex Financial Dictionary*. 1th. Farlex, Inc. URL: https://financial-dictionary.thefreedictionary.com/judgmental+forecast (visited on 04/18/2018).

Jung, Dominik and Verena Dorner (2018). "Decision Inertia and Arousal: Using NeuroIS to Analyze Bio-Physiological Correlates of Decision Inertia in a Dual-Choice Paradigm". In: *Information Systems and Neuroscience*. Ed. by Fred D Davis et al. Cham: Springer International Publishing, pp. 159–166. ISBN: 978-3-319-67431-5. DOI: 10.1007/978-3-319-67431-5_18.

Jurca, Radu and Boi Faltings (2008). "Incentives for Expressing Opinions in Online Polls". In: *Proceedings of the 9th ACM Conference on Electronic Commerce*. EC '08. New York, NY, USA: ACM, pp. 119–128. ISBN: 978-1-60558-169-9. DOI: 10.1145/1386790.1386812.

Kahneman, Daniel (2012). *Thinking, Fast and Slow*. Penguin. ISBN: 978-0141033570.

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler (1991). "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias". In: *The Journal of Economic Perspectives* 5.1, pp. 193–206. DOI: 10.1257/jep.5.1.193.

Kankanhalli, Atreyi, Bernard C Y Tan, and Kwok-Kee Wei (2005). "Contributing Knowledge to Electronic Knowledge Repositories: An Empirical Investigation". In: *MIS Quarterly* 29.1, pp. 113–143. ISSN: 02767783. DOI: 10.2307/25148670.

Karimi, Majid and Stanko Dimitrov (2018). "On the Road to Making Science of "Art": Risk Bias in Market Scoring Rules". In: *Decision Analysis*, pp. 1–18. DOI: 10.1287/deca.2017.0362Full.

Kaye, Anthony (1992). "Learning Together Apart". In: *Collaborative Learning Through Computer Conferencing*. Ed. by Anthony R Kaye. Berlin, Heidelberg: Springer. Chap. 1, pp. 1–24. ISBN: 978-3-642-77684-7. DOI: 10.1007/978-3-642-77684-7_1.

Keim, Donald B and Ananth Madhavan (1995). "Anatomy of the trading process empirical evidence on the behavior of institutional traders". In: *Journal of Financial Economics* 37.3, pp. 371–398. ISSN: 0304-405X.

Keller, Jonas and Heiko A. von der Gracht (2014). "The influence of information and communication technology (ICT) on future foresight processes - Results from a Delphi survey". In: *Technological Forecasting and Social Change* 85, pp. 81–92. ISSN: 00401625. DOI: 10.1016/j.techfore.2013.07.010.

Kennedy, Courtney et al. (2017). *An Evaluation of 2016 Election Polls in the U.S.* Tech. rep. American Association for Public Opinion Research. URL: https://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx.

Kim, Amy Jo (2000). *Community Building on the Web: Secret Strategies for Successful Online Communities*. 1st. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc. ISBN: 0201874849.

Kittur, Aniket et al. (2013). "The Future of Crowd Work". In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. CSCW '13. New York, NY, USA: ACM, pp. 1301–1318. ISBN: 978-1-4503-1331-5. DOI: 10.1145/2441776.2441923.

Klein, Mark and Ana Cristina Bicharra Garcia (2015). "High-speed idea filtering with the bag of lemons". In: *Decision Support Systems* 78, pp. 39–50. ISSN: 0167-9236. DOI: 10.1016/j.dss.2015.06.005.

Klingert, Frank M. A. and Matthias Meyer (2018). "Comparing Prediction Market Mechanisms: An Experiment-Based and Micro Validated Multi-Agent Simulation". In: *Journal of Artificial Societies and Social Simulation* 21.1, p. 7. ISSN: 1460-7425. DOI: 10.18564/jasss.3577.

Kloker, Simon (2016). "Application of the Dual-Process Theory to debias Forecasts in Prediction Markets". Accepted and presented at the HCOMP Doctorial Consortium 2016, Austin, US. URL: https://www.researchgate.net/publication/310241119_Application_of_the_Dual-Process_Theory_to_debias_Forecasts_in_Prediction_Markets.

Kloker, Simon and Tobias T Kranz (2017). "Manipulation in Prediction Markets - Chasing the Fraudsters". In: *Proceedings of the 25th European Conference of Information Systems, June 5th-10th 2017, Guimarães, Portugal*.

Kloker, Simon, Tim Straub, and Christof Weinhardt (2017a). "Designing a Crowd Forecasting Tool to Combine Prediction Markets and Real-Time Delphi". In: *Designing the Digital Transformation. DESRIST 2017. Lecture Notes in Computer Science*. Ed. by Alexander Maedche, Jan vom Brocke, and Alan Hevner. 10243rd ed. Springer, Cham: Springer International Publishing, pp. 468–473. ISBN: 978-3-319-59144-5. DOI: 10.1007/978-3-319-59144-5_33.

— (2017b). "Partition Dependence Bias in Prediction Markets". Accepted and presented at the Collective Intelligence Conference 2017, New York, US. URL: https://www.researchgate.net/publication/318126810_Partition_Dependence_Bias_in_Prediction_Markets.

— (N.D.). "The Influence of Partition Dependence in Forecasting Decisions (Working Paper)". unpublished.

Kloker, Simon et al. (2016). "Shouldn't Collaboration be social? – Proposal of a social Real-Time Delphi". In: *Proceedings of the Second Karlsruhe Service Summit Research Workshop*. URL: http://service-summit.ksri.kit.edu/downloads/Session_3B2_KSS_2016_paper_19.pdf.

Kloker, Simon et al. (2017). "Partition Dependence Bias in Group Forecasting". In: *Proceedings of the 17th International Conference on Group Decision and Negotiation*, pp. 9–19.

Kloker, Simon et al. (2018a). "Analyzing Prediction Market Trading Behavior to select Delphi-Experts". In: *Foresight* 20.4, pp. 364–374. DOI: 10.1108/FS-01-2018-0009.

Kloker, Simon et al. (2018b). "Fraud and Manipulation Prevention in Prediction Markets". In: *Proceedings of the 13th International Conference, DESRIST 2018, Chennai, India, June 3–6, 2018*, pp. 1–6.

— (2018c). "The Effect of Social Reputation on Retention: Designing a Social Real-Time Delphi Platform". In: *Proceedings of the 26th European Conference on Information Systems (ECIS2018), Portsmouth,UK, 2018*.

Kloker, Simon et al. (2019). "Delphi-Märkte". In: *Delphi-Verfahren: Konzept, Varianten und Einsatzbereiche in der Gesundheitswissenschaft*. Ed. by Marlen Niederberger and Ortwin Renn. Springer Berlin Heidelberg. Chap. 5, forthcoming.

Kloker, Simon et al. (N.D.). "Crowd-sourced Manipulation and Fraud Detection: An Action Design Research Study in Prediction Markets (Working Paper)". unpublished.

Knöll, Florian and Viliam Simko (2017). "Organizational Information improves Forecast Efficiency of Correction Techniques". In: *Proceedings of the 17th Conference on Information Technologies - Applications and Theory (ITAT), CEUR Workshop*. Vol. 1885, pp. 86–92.

Kochtanek, Thomas R and Karen K Hein (1999). "Delphi study of digital libraries". In: *Information Processing & Management* 35.3, pp. 245–254. ISSN: 0306-4573. DOI: 10.1016/S0306-4573(98)00060-0.

Kramer Mayer, Marcia and Paul J. Hinton (2010). *How Crowdsourcing Could Help the SEC*. URL: https://hbr.org/2010/08/how-crowdsourcing-could-help-t (visited on 10/13/2017).

Kranz, Tobias T (2015). "Continuous Market Engineering - Focusing Agent Behaviour, Interfaces, and Auxiliary Services". PhD thesis. Universität Karlsruhe.

Kranz, Tobias T, Florian Teschner, and Christof Weinhardt (2014). "Combining Prediction Markets and Surveys: An Experimental Study". In: *Proceedings of the European Conference on Information Systems (ECIS) 2014, Tel Aviv, Israel, June 9-11, 2014*.

Kranz, Tobias T et al. (2014). "Identifying Individual Party Preferences in Political Stock Markets". In: *Proceedings of the IADIS International Conference on E-Society. (Madrid, Spain)*, pp. 162–169.

Kuechler, Bill and Vijay Vaishnavi (2008). "On theory development in design science research: anatomy of a research project". In: *European Journal of Information Systems* 17.5, pp. 489–504. ISSN: 1476-9344. DOI: 10.1057/ejis.2008.40.

Kuusi, Osmo (1999). *Expertise in the Future Use of Generic Technologies*. Tech. rep. VATT Institute for Economic Research.

Kyle, Albert S (2016). "Continuous Auctions and Insider Trading". In: *Econometrica* 53.6, pp. 1315–1335. DOI: 10.3982/ECTA6822.

LaComb, Christina Ann, Janet Arlie Barnett, and Qimei Pan (2007). "The imagination market". In: *Information Systems Frontiers* 9.2, pp. 245–256. ISSN: 1572-9419. DOI: 10.1007/s10796-007-9024-9.

Lampe, Cliff et al. (2010). "Motivations to Participate in Online Communities". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. New York, NY, USA: ACM, pp. 1927–1936. ISBN: 978-1-60558-929-9. DOI: 10.1145/1753326.1753616.

Lampel, Joseph and Ajay Bhalla (2007). "The Role of Status Seeking in Online Communities: Giving the Gift of Experience". In: *Journal of Computer-Mediated Communication* 12.2, pp. 434–455. ISSN: 1083-6101. DOI: 10.1111/j.1083-6101.2007.00332.x.

Landeta, Jon (2006). "Current validity of the Delphi method in social sciences". In: *Technological Forecasting and Social Change* 73.5, pp. 467–482. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2005.09.002.

Laskey, Kathryn Blackmond, Robin Hanson, and C Twardy (2015). "Combinatorial prediction markets for fusing information from distributed experts and models". In: *Proceedings of the 18th International Conference on Information Fusion (Fusion), 2015*, pp. 1892–1898.

Lawrence, Michael et al. (2006). "Judgmental forecasting: A review of progress over the last 25 years". In: *International Journal of Forecasting* 22.3, pp. 493–518. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2006.03.007.

Lee, Jae Kook et al. (2014). "Social Media, Network Heterogeneity, and Opinion Polarization". In: *Journal of Communication* 64.4, pp. 702–722. ISSN: 1460-2466. DOI: 10.1111/jcom.12077.

Leuthold, Raymond M, Philip Garcia, and Richard Lu (1994). "The returns and forecasting ability of large traders in the frozen pork bellies futures market". In: *The Journal of Business* 67.3, pp. 459–473. ISSN: 0021-9398.

Levin, Irwin P, Daniel P Chapman, and Richard D Johnson (1988). "Confidence in judgments based on incomplete information: An investigation using both hypothetical and real gambles". In: *Journal of Behavioral Decision Making* 1.1, pp. 29–41. ISSN: 1099-0771. DOI: 10.1002/bdm.3960010105.

Li, Xiaolong and Jennifer Wortman Vaughan (2013). "An Axiomatic Characterization of Adaptive-liquidity Market Makers". In: *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*. EC '13. New York, NY, USA: ACM, pp. 657–674. ISBN: 978-1-4503-1962-1. DOI: 10.1145/2482540.2482575.

Lin, Guan-Yu (2004). "Social Presence Questionnaire of Online Collaborative Learning: Development and Validity". In: *Association for Educational Communications and Technology, 27th, Chicago, IL, October 19-23, 2004*, pp. 588–591. URL: http://files.eric.ed.gov/fulltext/ED484999.pdf.

Linstone, Harold A and Murray Turoff (2002a). "General Applications". In: *The Delphi Method: Techniques and applications (edited version of the book from 1975)*. Ed. by Harold A Linstone and Murray Turoff. Chap. 3, pp. 69–220.

— (2002b). "Introduction". In: *The Delphi Method: Techniques and applications (edited version of the book from 1975)*. Ed. by Harold A Linstone and Murray Turoff. Chap. 1, pp. 1–12.

— (2002c). *The Delphi Method: Techniques and applications*. Addison-Wesley Pub. Co.

— (2011). "Delphi: A brief look backward and forward". In: *Technological Forecasting and Social Change* 78.9, pp. 1712–1719. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2010.09.011.

Lou, Y. and M. Wang (2011). "Fraud Risk Factor of the Fraud Triangle Assessing the Likelihood of Fraudulent Financial Reporting". In: *Journal of Business & Economics Research (JBER)* 7.2, pp. 61–79. DOI: 10.19030/jber.v7i2.2262.

Luckner, Stefan (2006). "Prediction Markets: How Do Incentive Schemes Affect Prediction Accuracy?" In: *Negotiation and Market Engineering'06*, pp. 1–10.

Luckner, Stefan, Felix Kratzer, and Christof Weinhardt (2005). "STOCCER-A Forecasting Market for the FIFA World Cup 2006". In: *4th Workshop on e-Business (WeB 2005), Las Vegas, USA*.

Luckner, Stefan and Christof Weinhardt (2007). "How to Pay Traders in Information Markets: Results from a Field Experiment". In: *Journal of Prediction Markets* 1.2, pp. 147–156. URL: http://econpapers.repec.org/RePEc:buc:jpredm:v:1:y:2007:i:2:p:147-156.

— (2008). "Arbitrage opportunities and market-making traders in prediction markets". In: *Proceedings of the 10th IEEE Conference on E-Commerce Technology and the 5th IEEE Conference on Enterprise Computing, E-Commerce and E-Services, 2008*. IEEE, pp. 53–59. ISBN: 076953340X.

Luckner, Stefan et al. (2012). *Prediction markets: Fundamentals, designs, and applications*. Springer Science & Business Media. ISBN: 3834970859.

Luskin, Donald L (2004). *Who'S Behind The Bush-Futures Attacks?* URL: http://www.nationalreview.com/article/212580/whos-behind-bush-futures-attacks-donald-l-luskin (visited on 04/06/2017).

Lyon, Aidan and Eric Pacuit (2013). "The Wisdom of Crowds: Methods of Human Judgement Aggregation". In: *Handbook of Human Computation*. Ed. by Pietro Michelucci. Springer New York. Chap. 5.3, pp. 599–614. ISBN: 978-1-4614-8806-4. DOI: 10.1007/978-1-4614-8806-4_47.

Malekovic, Ninoslav, Juliana Sutanto, and Lazaros Goutas (2016). "Manipulative Imputation in Distributed Decision Support Settings: The Implications of Information Asymmetry and Aggregation Complexity". In: *Decision Support Systems* 85, pp. 1–11. ISSN: 0167-9236. DOI: 10.1016/j.dss.2016.02.004.

Malkiel, Burton G (2003). "The Efficient Market Hypothesis and Its Critics". In: *The Journal of Economic Perspectives* 17.1, pp. 59–82. ISSN: 08953309. URL: http://www.jstor.org/stable/3216840.

Mangold, B. et al. (2005). "The Tech Buzz Game: stock market prediction". In: *Computer* 38.7, pp. 94–97. DOI: 10.1109/MC.2005.243.

Mason, Winter and Siddharth Suri (2012). "Conducting behavioral research on Amazon's Mechanical Turk". In: *Behavior Research Methods* 44.1, pp. 1–23. ISSN: 1554351X. DOI: 10.3758/s13428-011-0124-6. arXiv: /ssrn.com/abstract=1691163 [http:].

Matti, Thimothy, Yuntao Zhu, and Kuai Xu (2014). "Financial fraud detection using social media crowdsourcing". In: *Proceedings of the IEEE 33rd International Performance Computing and Communications Conference (IPCCC) 2014*, pp. 1–2. DOI: 10.1109/PCCC.2014.7017023.

McFadden, Daniel (1973). *Conditional logit analysis of qualitative choice behavior*. DOI: 10.1108/eb028592.

Menkhoff, Lukas and Maik Schmeling (2010). "Whose trades convey information? Evidence from a cross-section of traders". In: *Journal of Financial Markets* 13.1, pp. 101–128. ISSN: 1386-4181.

Meservy, Thomas O, Matthew L Jensen, and Kelly J Fadel (2013). "Evaluation of Competing Candidate Solutions in Electronic Networks of Practice". In: *Information Systems Research* 25.1, pp. 15–34. ISSN: 1047-7047. DOI: 10.1287/isre.2013.0502.

Meub, Lukas and Till Proeger (2016). "Can anchoring explain biased forecasts? Experimental evidence". In: *Journal of Behavioral and Experimental Finance* 12, pp. 1–13. ISSN: 22146369. DOI: 10.1016/j.jbef.2016.08.001.

Miller, Michael K et al. (2012). "Citizen Forecasts of the 2008 U.S. Presidential Election". In: *Politics & Policy* 40.6, pp. 1019–1052. ISSN: 1747-1346. DOI: 10.1111/j.1747-1346.2012.00394.x.

Moore, Tyler and Richard Clayton (2008). "Evaluating the wisdom of crowds in assessing phishing websites". In: *Financial Cryptography and Data Security: 12th International Conference, FC 2008, Cozumel, Mexico, January 28-31, 2008. Revised Selected Papers*. Ed. by Gene Tsudik. Vol. 5143. Springer. Chap. 2, pp. 16–30.

Mullen, Penelope M (2003). "Delphi: myths and reality". In: *Journal of Health Organization and Management* 17.1, pp. 37–52. DOI: 10.1108/14777260310469319.

Müller-Trede, Johannes et al. (2018). "The Wisdom of Crowds in Matters of Taste". In: *Management Science* 64.4, pp. 1779–1803. ISSN: 0025-1909. DOI: 10.1287/mnsc.2016.2660.

Munafo, Marcus R et al. (2015). "Using prediction markets to forecast research evaluations". In: *Royal Society Open Science* 2.10, pp. 1–8. DOI: 10.1098/rsos.150287.

Murphy, Elizabeth (2004). "Recognising and promoting collaboration in an online asynchronous discussion". In: *British Journal of Educational Technology* 35.4, pp. 421–431. URL: http://www.ucs.mun.ca/~emurphy/bjet_401.pdf.

Nagar, Yiftach and Thomas W Malone (2012). "Improving Predictions with Hybrid Markets". In: *Proceedings of the American Association of Artificial Intelligence (AAAI) Fall Symposium on Machine Aggregation of Human Judgment, Arlington, VA, November 2-4, 2012*, pp. 30–36.

Nakamura, Jeanne and Mihaly Csikszentmihalyi (2014). "The Concept of Flow". In: *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*. Ed. by Mihaly Csikszentmihalyi. Dordrecht: Springer Netherlands, pp. 239–263. ISBN: 978-94-017-9088-8. DOI: 10.1007/978-94-017-9088-8_16.

Newman, Alexandra Lee (2010). "Manipulation in Political Prediction Markets". In: *Journal of Business, Entrepreneurship & the Law* 3.2, pp. 205–235. URL: http://digitalcommons.pepperdine.edu/jbel/vol3/iss2/1/.

Nielsen, J (2006). *The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities*. URL: https://www.nngroup.com/articles/participation-inequality/ (visited on 02/12/2018).

Niemeyer, Claudia et al. (2016). "Participatory crowdfunding: An approach towards engaging employees and citizens in institutional budgeting decisions". In: *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 2800–2808. ISBN: 9780769556703. DOI: 10.1109/HICSS.2016.351.

Norman, Donald A (1998). *The Design of Everyday Things*. Currency Doubleday, New York.

O'Connor, Philip and Feng Zhou (2008). "The Tradesports NFL Prediction Market: An Analysis of Market Efficiency, Transaction Costs, and Bettor Preferences". In: *Journal of Prediction Markets* 2.1, pp. 45–71.

Okoli, Chitu and Suzanne D Pawlowski (2004). "The Delphi method as a research tool: an example, design considerations and applications". In: *Information & Management* 42.1, pp. 15–29. ISSN: 0378-7206. DOI: 10.1016/j.im.2003.11.002.

Oliven, Kenneth and Thomas A Rietz (2004). "Suckers are born but markets are made: Individual rationality, arbitrage, and market efficiency on an electronic futures market". In: *Management Science* 50.3, pp. 336–351. ISSN: 0025-1909.

Oprea, Ryan et al. (2008). "Can Manipulators Mislead Prediction Market Observers?"

Othman, Abraham and Tuomas Sandholm (2010). "Decision Rules and Decision Markets". In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*. Toronto, Canada: International Foundation for Autonomous Agents and Multiagent Systems, pp. 625–632. URL: http://dl.acm.org/citation.cfm?id=1838206.1838288.

Ottaviani, Marco and Peter Norman Sørensen (2007). "Outcome manipulation in corporate prediction markets". In: *Journal of the European Economic Association* 5.2-3, pp. 554–563. ISSN: 1542-4774. DOI: 10.1162/jeea.2007.5.2-3.554.

Page, Lionel (2012). "'It ain't over till it's over.'Yogi Berra bias on prediction markets". In: *Applied Economics* 44.1, pp. 81–92.

Pennock, David M (2010). *Why automated market makers ? | Oddhead Blog*. URL: http://blog.oddhead.com/2010/07/08/why-automated-market-makers/ (visited on 05/14/2018).

Pennock, David M et al. (2001). "The Power of Play: Efficiency and Forecast Accuracy in Web Market Games". In: *Science* 291, pp. 987–988. URL: http://artificial-markets.com/am/pennock-neci-tr-2000-168/.

Peterson, Jack et al. (2018). *Augur: a Decentralized Oracle and Prediction Market Platform*. Tech. rep. Forecast Foundation. arXiv: 1501.01042. URL: https://www.augur.net/whitepaper.pdf.

Pouget, Sebastien, Julien Sauvagnat, and Stephane Villeneuve (2017). "A mind is a terrible thing to change: confirmatory bias in financial markets". In: *The Review of Financial Studies* 30.6, pp. 2066–1209. DOI: 10.1093/rfs/hhw100.

Powell, Walter A. et al. (2013). "Combinatorial prediction markets: An experimental study". In: *Scalable Uncertainty Management*. Vol. 8078 LNAI. Springer Berlin Heidelberg, pp. 283–296. ISBN: 9783642403804. DOI: 10.1007/978-3-642-40381-1-22.

Prelec, Dražen (2004). "A Bayesian Truth Serum for Subjective Data". In: *Science* 306.5695, pp. 462–467. DOI: 10.1126/science.1102081.

Prokesch, Tobias, Heiko A von der Gracht, and Holger Wohlenberg (2015). "Integrating prediction market and Delphi methodology into a foresight support system — Insights from an online game". In: *Technological Forecasting and Social Change* 97, pp. 47–64. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2014.02.021.

Qiu, Liangfei and Subodha Kumar (2017). "Understanding Voluntary Knowledge Provision and Content Contribution Through a Social-Media-Based Prediction Market: A Field Experiment". In: *Information Systems Research* 28.3, pp. 529–546. ISSN: 1047-7047. DOI: 10.1287/isre.2016.0679.

Rainie, Lee (2007). *28% of online americans have used the internet to tag content*. Tech. rep. URL: http://www.pewinternet.org/files/old-media/Files/Reports/2007/PIP_Tagging.pdf.

Reichelson, Sheri et al. (2017). "Partition dependence in consumer choice: Perceptual groupings do not reliably shape decisions". In: *Psychonomic Bulletin & Review*. ISSN: 1531-5320. DOI: 10.3758/s13423-017-1326-4.

Reid, N (1988). "The Delphi technique: Its contribution to the evaluation of professional practice". In: *Professional Competence and Quality Assurance in the Caring Professions*. Ed. by Roger Ellis. Chapman & Hall, London. Chap. 9, pp. 230–254. DOI: 10.1016/0020-7489(90)90106-S.

Reips, Ulf-Dietrich (2007). "The methodology of Internet-based experiments". In: *The Oxford handbook of Internet psychology*. Ed. by Adam Joinson et al. Oxford University Press Oxford, UK. Chap. 24, pp. 373–390. ISBN: 9780198568001.

Remidez Jr, Herbert and Curtis Joslin (2007). "Using Prediction Markets to Support IT Project Management - Research in Progress". In: *International Research Workshop on IT Project Management 2007*, p. 10.

Restocchi, Valerio et al. (2018). "It takes all sorts: A heterogeneous agent explanation for prediction market mispricing". In: *European Journal of Operational Research* In Press, pp. 1–14. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2018.04.011.

Reuters, Editorial (2010). *Hong Kong warrants traders guilty of market manipulation: SFC*. URL: http://www.reuters.com/article/us-hongkong-sfc-idUSTRE6462\-GN20100507 (visited on 04/06/2017).

Rhode, Paul W and Koleman S Strumpf (2004). "Historical Presidential Betting Markets". In: *Journal of Economic Perspectives* 18.2, pp. 127–141. URL: http://www.aeaweb.org/articles?id=10.1257/0895330041371277.

— (2008). "Manipulating Political Stock Markets: A Field Experiment and a Century of Observational Data". URL: https://www.unc.edu/~cigar/papers/ManipIHT_June2008(KS).pdf.

Rodrigues, Eduarda Mendes, Natasa Milic-Frayling, and Blaz Fortuna (2008). "Social Tagging Behaviour in Community-Driven Question Answering". In: *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*. WI-IAT '08. Washington, DC, USA: IEEE Computer Society, pp. 112–119. ISBN: 978-0-7695-3496-1. DOI: 10.1109/WIIAT.2008.138.

Rohrbeck, René, Nico Thom, and Heinrich Arnold (2015). "IT tools for foresight: The integrated insight and response system of Deutsche Telekom Innovation Laboratories". In: *Technological Forecasting and Social Change* 97, pp. 115–126. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2013.09.015.

Roschelle, Jeremy and Stephanie D Teasley (1995). "The Construction of Shared Knowledge in Collaborative Problem Solving". In: *Computer Supported Collaborative Learning*. Ed. by Claire O'Malley. Vol. 128. NATO ASI Series. Springer Berlin Heidelberg. Chap. 1,5, pp. 69–97. DOI: 10.1007/978-3-642-85098-1_5.

Rosenbloom, Earl S and William Notz (2006). "Statistical Tests of Real Money versus Play Money Prediction Markets". In: *Electronic Markets* 16.1, pp. 63–69. ISSN: 1019-6781.

Roth, Alvin E. (1986). *Laboratory experimentation in economics*. Vol. 2. 2. Cambridge: Cambridge University Press, pp. 245–273. ISBN: 9780511528316. DOI: 10.1017/S1478061500002656.

Rothschild, David (2009). "Forecasting Elections: Comparing Prediction Markets, Polls, and Their Biases". In: *Public Opinion Quarterly* 73.5, pp. 895–916. DOI: 10.1093/poq/nfp082. URL: http://poq.oxfordjournals.org/content/73/5/895.abstract.

Rothschild, David and Rajiv Sethi (2016). "Trading Strategies and Market Microstructure: Evidence from a Prediction Market." In: *Journal of Prediction Markets* 10.1, pp. 1–29. ISSN: 17506751. URL: http://www.redi-bw.de/db/ebsco.php/search.ebscohost.com/login.aspx?direct=true&db=buh&AN=118713442&site=eds-live.

Rothschild, David and Justin Wolfers (2013). "Forecasting Elections: Voter Intentions Versus Expectations". URL: https://ssrn.com/abstract=1884644.

Rovai, Alfred P (2001). "Building classroom community at a distance: A case study". In: *Educational Technology Research and Development* 49.4, p. 33. ISSN: 1556-6501. DOI: 10.1007/BF02504946.

Rowe, Dawn A. et al. (2015). "A Delphi study to operationalize evidence-based predictors in secondary transition". In: *Career Development and Transition for Exceptional Individuals* 38.2, pp. 113–126. ISSN: 2165-1434. DOI: 10.1177/21651434145\-26429.

Rowe, Gene and George Wright (1999). "The Delphi technique as a forecasting tool: issues and analysis". In: *International Journal of Forecasting* 15.4, pp. 353–375. ISSN: 0169-2070. DOI: 10.1016/S0169-2070(99)00018-7.

Rowe, Gene, George Wright, and Fergus Bolger (1991). "Delphi: A reevaluation of research and theory". In: *Technological Forecasting and Social Change* 39.3, pp. 235–251. ISSN: 00401625. DOI: 10.1016/0040-1625(91)90039-I.

Rowe, Gene, George Wright, and Andy McColl (2005). "Judgment change during Delphi-like procedures: The role of majority influence, expertise, and confidence". In: *Technological Forecasting and Social Change* 72.4, pp. 377–399. ISSN: 00401625. DOI: 10.1016/j.techfore.2004.03.004.

Samson, Alain and Benjamin G Voyer (2012). "Two minds, three ways: dual system and dual process models in consumer psychology". In: *AMS Review* 2.2-4, pp. 48–71.

Schlag, Karl H, James Tremewan, and Joël J van der Weele (2015). "A penny for your thoughts: a survey of methods for eliciting beliefs". In: *Experimental Economics* 18.3, pp. 457–490. ISSN: 1573-6938. DOI: 10.1007/s10683-014-9416-x.

Schröder, Jan (2009). *Manipulations in prediction markets: analysis of trading behaviour not conforming with trading regulations*. KIT Scientific Publishing. ISBN: 3866443447.

Schuckmann, Steffen W. et al. (2012). "Analysis of factors influencing the development of transport infrastructure until the year 2030 — A Delphi based scenario study". In: *Technological Forecasting and Social Change* 79.8, pp. 1373–1387. ISSN: 00401625. DOI: 10.1016/j.techfore.2012.05.008.

Sein, Maung K et al. (2011). "Action Design Research". In: *MIS Quarterly* 35.1, pp. 37–56. ISSN: 02767783. DOI: 10.2307/23043488. URL: http://www.jstor.org/stable/23043488.

Servan-Schreiber, Emile (2017). "Debunking Three Myths About Crowd - Based Forecasting". In: *Collective Intelligence Conference*.

Servan-Schreiber, Emile et al. (2004). "Prediction Markets: Does Money Matter?" In: *Electronic Markets* 14.3, pp. 243–251. DOI: 10.1080/1019678042000245254.

Seuken, Sven (2010). "Hidden Market Design". In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*. AAMAS '10. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1661–1662. ISBN: 978-0-9826571-1-9.

Sewell, Martin (2011). "History of the efficient market hypothesis". URL: https://www.gyc.com.sg/files/p_sewell-history.pdf.

Sharifi, M, E Fink, and J G Carbonell (2011). "SmartNotes: Application of crowdsourcing to the detection of web threats". In: *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1346–1350. DOI: 10.1109/ICSMC.2011.6083845.

Shayo, Deodatus Patrick and Norbert Kersting (2017). "Crowdmonitoring of Elections through ICT: The Case of Uchaguzi Wetu 2015 Crowdsourcing Platform in Tanzania". In: *Proceedings of the 2017 Conference for E-Democracy and Open Government (CeDEM)*, pp. 36–45. DOI: 10.1109/CeDEM.2017.13.

Shrier, David et al. (2016). *Prediction Markets*. Tech. rep. Massachusetts Institute of Technology, p. 18. URL: http://cdn.resources.getsmarter.ac/wp-content/uploads/2016/08/mit_prediction_markets_report.pdf.

Skinner, David C (2009). *Introduction to Decision Analysis*. 3rd ed. Probabilistic Publishing, p. 368. ISBN: 978-0964793866.

Slamka, Christian, Wolfgang Jank, and Bernd Skiera (2012). "Second-Generation Prediction Markets for Information Aggregation: A Comparison of Payoff Mechanisms". In: *Journal of Forecasting* 31, pp. 469–489. DOI: 10.1002/for.1225.

Slamka, Christian, Arina Soukhoroukova, and Martin Spann (2008). "Event studies in real-and play-money prediction markets". In: *Journal of Prediction Markets* 2.2, pp. 53–70. ISSN: 1750-6751.

Sniezek, Janet A (1990). "A Comparison of Techniques for Judgmental Forecasting by Groups with Common Information". In: *Group & Organization Studies* 15.1, pp. 5–19. ISSN: 0364-1082. DOI: 10.1177/105960119001500102.

Snowberg, Erik and Justin Wolfers (2010). "Explaining the Favorite-Longshot Bias: Is it Risk-Love or Misperceptions?" In: *Journal of Political Economy* 118.4, pp. 723–746. DOI: 10.1086/655844.

Snowberg, Erik, Justin Wolfers, and Eric Zitzewitz (2005). "Information (in)efficiency in prediction markets". In: *Information Efficiency in Financial and Betting Markets*. Leighton Vaughan Williams. Chap. 18, pp. 266–386. ISBN: 9780511493614. DOI: 10.1017/CBO9780511493614.019.

Sobel, Russell S. and S. Travis Raines (2003). "An examination of the empirical derivatives of the favourite-longshot bias in racetrack betting". In: *Applied Economics* 35.4, pp. 371–385. ISSN: 00036846. DOI: 10.1080/00036840110111176.

Sonnemann, Ulrich et al. (2011). "Partition Dependence in Prediction Markets: Field and Lab Evidence".

— (2013). "How psychological framing affects economic market prices in the lab and field". In: *Proceedings of the National Academy of Sciences* 110.29, pp. 11779–11784.

Soukhoroukova, Arina and Martin Spann (2005). "New Product Development with Internet Based Information Markets: Theory and Empirical Application". In: *Proceedings of the 13th European Conference on Information Systems (ECIS), Regensburg*. URL: http://aisel.aisnet.org/ecis2005/133.

Soukhoroukova, Arina, Martin Spann, and Bernd Skiera (2012). "Sourcing, filtering, and evaluating new product ideas: An empirical exploration of the performance of idea markets". In: *Journal of Product Innovation Management* 29.1, pp. 100–112. ISSN: 1540-5885.

Spann, Martin and Bernd Skiera (2009). "Sports Forecasting : A Comparison of the Forecast Accuracy of Prediction Markets , Betting Odds and Tipsters Sports Forecasting : A Comparison of the Forecast Accuracy of Prediction Markets , Betting Odds and Tipsters". In: *Journal of Forecasting* 28, pp. 55–72. DOI: 10.1002/for.1091.

Spears, Brian and Christina LaComb (2009). "Examining trader behavior in idea markets: an implementation of GE's imagination markets". In: *The Journal of Prediction Markets* 3.1, pp. 17–39. ISSN: 1750-676X.

Sprenger, Timm, Paul Bolster, and Anad Venkateswaran (2007). "Conditional Prediction Markets as Corporate Decision Support Systems - An Experimental Comparison with Group Deliberation". In: *Journal of Prediction Markets* 1.3, pp. 189–208. DOI: 10.5750/jpm.v1i3.428.

Stastny, Bradley J and Paul E Lehner (2018). "Comparative evaluation of the forecast accuracy of analysis reports and a prediction market". In: *Judgment and Decision Making* 13.2, pp. 202–211.

Steinert, Martin (2009). "A dissensus based online Delphi approach: An explorative research tool". In: *Technological Forecasting and Social Change* 76.3, pp. 291–300. ISSN: 00401625. DOI: 10.1016/j.techfore.2008.10.006.

Straub, Tim, Timm Teubner, and Christof Weinhardt (2016). "Risk Taking in Online Crowdsourcing Tournaments". In: *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 1851–1860. DOI: 10.1109/HICSS.2016.235.

Styhre, Alexander (2002). "The knowledge-intensive company and the economy of sharing: rethinking utility and knowledge management". In: *Knowledge and Process Management* 9.4, pp. 228–236. ISSN: 1099-1441. DOI: 10.1002/kpm.155.

Sun, Wei et al. (2012). "Probability and Asset Updating using Bayesian Networks for Combinatorial Prediction Markets". In: *Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 815–824. ISBN: 9780974903989. arXiv: 1210.4900. URL: http://arxiv.org/abs/1210.4900.

Tannenbaum, David et al. (2015). "Nudging Physician Prescription Decisions by Partitioning the Order Set: Results of a Vignette-Based Study". In: *Journal of General Internal Medicine* 30.3, pp. 298–304. ISSN: 1525-1497. DOI: 10.1007/s11606-014-3051-2.

Teschner, Florian (2011). "Behavioral ICT: Risk, Cognition and Information". In: *Proceedings of the Doctoral Consortium, Wirtschaftsinformatik 2011*. Zurich, Switzerland.

— (2012). *Forecasting Economic Indices: Design, Performance, and Learning in Prediction Markets*. Karlsruhe Institute of Technology (KIT). URL: http://digbib.ubka.uni-karlsruhe.de/volltexte/1000029512.

Teschner, Florian, Stephan Stathel, and Christof Weinhardt (2011). "A prediction market for macro-economic variables". In: *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 1–9. ISBN: 9780769542829. DOI: 10.1109/HICSS.2011.23.

Teschner, Florian and Christof Weinhardt (2012a). "Evaluating Hidden Market Design". In: *Auctions, Market Mechanisms, and Their Applications*. Ed. by Peter Coles et al. Vol. 80. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer Berlin Heidelberg, pp. 5–17. ISBN: 978-3-642-30912-0. DOI: 10.1007/978-3-642-30913-7_3.

— (2012b). "Identifying Experts in Virtual Forecasting Communities". In: *AMCIS 2012 Proceedings*. URL: http://aisel.aisnet.org/amcis2012/proceedings/VirtualCommunities/4.

Tetlock, Philip E and Dan Gardner (2015). *Superforecasting: The Art and Science of Prediction*. 1st ed. Crown, p. 352. ISBN: 978-0804136693.

Tetlock, Philip E, Barbara A Mellers, and J Peter Scoblic (2017). "Bringing probability judgments into policy debates via forecasting tournaments". In: *Science* 355.6324, 481 LP –483. URL: http://science.sciencemag.org/content/355/6324/481.abstract.

Teubner, Timm et al. (2013). "Social identity and reciprocity in online gift giving networks". In: *Proceedings of the 46th Annual Hawaii International Conference on System Sciences*, pp. 708–717. ISBN: 9780769548920. DOI: 10.1109/HICSS.2013.489.

Theodoulidis, Babis and David Diaz (2012). "Financial Markets and High Frequency Trading: An Information Management Perspective". URL: https://ssrn.com/abstract=2178944.

Trumbo, Craig W (2002). "Information Processing and Risk Perception: An Adaptation of the Heuristic-Systematic Model". In: *Journal of Communication* 52.2, pp. 367–382. ISSN: 1460-2466. DOI: 10.1111/j.1460-2466.2002.tb02550.x.

Turoff, Murray et al. (2004). "Online Collaborative Learning Enhancement Through the Delphi Method". In: *Proceedings of the OZCHI 2004 Conference, University of Wollongong, Australia, November 22-24, 2004.* Pp. 66–79. URL: https://web.njit.edu/~turoff/Papers/ozchi2004.htm.

Tversky, Amos and Daniel Kahneman (1974). "Judgment under uncertainty: Heuristics and biases". In: *Science* 185.4157, pp. 1124–1131. ISSN: 0036-8075.

— (1983). "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment". In: *Psychological Review* 90.4, pp. 293–315. DOI: 10.1.1.304.6549.

Unerman, Jeffrey and Brendan O'Dwyer (2004). "Enron, WorldCom, Andersen et al.: a challenge to modernity". In: *Critical Perspectives on Accounting* 15.6–7, pp. 971–993. ISSN: 1045-2354. DOI: 10.1016/j.cpa.2003.04.002.

Vakkari, Pertti (1999). "Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval". In: *Information Processing & Management* 35.6, pp. 819–837. ISSN: 03064573. DOI: 10.1016/S0306-4573(99)00028-X.

Van Bruggen, Gerrit H. et al. (2010). "Prediction Markets as institutional forecasting support systems". In: *Decision Support Systems* 49.4, pp. 404–416. ISSN: 01679236. DOI: 10.1016/j.dss.2010.05.002.

Veiga, Helena and Marc Vorsatz (2010). "Information aggregation in experimental asset markets in the presence of a manipulator". In: *Experimental Economics* 13.4, pp. 379–398. ISSN: 1573-6938. DOI: 10.1007/s10683-010-9247-3.

Venable, John, Jan Pries-Heje, and Richard Baskerville (2016). "FEDS: a Framework for Evaluation in Design Science Research". In: *European Journal of Information Systems* 25.1, pp. 77–89. ISSN: 1476-9344. DOI: 10.1057/ejis.2014.36.

Vernon, Wesley (2009). "The Delphi technique: A review". In: *International Journal of Therapy and Rehabilitation* 16.2, pp. 69–76. ISSN: 1741-1645. DOI: 10.12968/ijtr.2009.16.2.38892.

Vila, Jean-Luc (1989). "Simple games of market manipulation". In: *Economics Letters* 29.1, pp. 21–26. ISSN: 0165-1765.

Voigt, Matthias, Björn Niehaves, and Jörg Becker (2012). "Towards a Unified Design Theory for Creativity Support Systems". In: *Design Science Research in Information Systems. Advances in Theory and Practice*. Ed. by Ken Peffers, Marcus Rothenberger, and Bill Kuechler. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 152–173. ISBN: 978-3-642-29863-9.

Wagenknecht, Thomas, René Filpe, and Christof Weinhardt (2017). "Towards a design theory of computer-supported organizational participation". In: *Journal of Enterprise Information Management* 30.1, pp. 188–202. ISSN: 1741-0398. DOI: 10.1108/JEIM-01-2016-0007.

Wagner, Christian and Andrea Back (2008). "Group Wisdom Support Systems: Aggregating the Insights of many trough Information Technology". In: *Issues in Information Systems (IIS)* 9.2, pp. 343–350. URL: http://iacis.org/iis/2008/S2008_992.pdf.

Walker, A M and J Selfe (1996). "The Delphi method: a useful tool for the allied health researcher". In: *British Journal of Therapy and Rehabilitation* 3.12, pp. 677–681.

Wang, Gang et al. (2013). "Social Turing Tests: Crowdsourcing Sybil Detection". In: *Proc. of The 20th Annual Network Distributed System Security Symposium (NDSS)*. URL: http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2017/09/06_1_0.pdf.

Wang, Hsiu-Yuan and Shwu-Huey Wang (2010). "Predicting mobile hotel reservation adoption: Insight from a perceived value standpoint". In: *International Journal of Hospitality Management* 29.4, pp. 598–608. ISSN: 0278-4319. DOI: 10.1016/j.ijhm.2009.11.001.

Watts, Stephanie (2015). "Application of Dual-process Theory to Information Systems: Current and Future Research Directions". In: *Foundations and Trends® in Information Systems*. Foundations and Trends(R) in Information Systems 1.2, pp. 69–162. DOI: 10.1561/2900000004.

Watts, Stephanie, G Shankaranarayanan, and Adir Even (2009). "Data quality assessment in context: A cognitive perspective". In: *Decision Support Systems* 48.1, pp. 202–211. ISSN: 0167-9236. DOI: 10.1016/j.dss.2009.07.012.

Webster, Jane and Richard T Watson (2002). "Analyzing the Past to Prepare for the Future: Writing a Literature Review." In: *MIS Quarterly* 26.2, pp. xiii –xxiii. ISSN: 02767783. DOI: 10.1.1.104.6570. arXiv: 02767783.

Weinhardt, Christof, Carsten Holtmann, and Dirk Neumann (2003). "WI-Schlagwort: Market-Engineering". In: *Wirtschaftsinformatik* 45.6, pp. 635–640.

Welty, Gordon (1972). "Problems of selecting experts for Delphi exercises". In: *Academy of Management Journal* 15.1, pp. 121–124. ISSN: 0001-4273.

Wikipedia (2018). *Frankfurter Allgemeine Zeitung — Wikipedia, The Free Encyclopedia*. http://de.wikipedia.org/w/index.php?title=Frankfurter%20Allgemeine%20Zeitung&oldid=176592654. [Online; accessed 26-April-2018].

Wilde, Tim R W de, Femke S Ten Velden, and Carsten K W De Dreu (2018). "The anchoring-bias in groups". In: *Journal of Experimental Social Psychology* 76, pp. 116–126. ISSN: 0022-1031. DOI: 10.1016/j.jesp.2018.02.001.

Winkler, Jens and Roger Moser (2016). "Biases in future-oriented Delphi studies: A cognitive perspective". In: *Technological Forecasting and Social Change* 105, pp. 63–76. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2016.01.021.

Wirth, Rüdiger and Jochen Hipp (2000). "CRISP-DM : Towards a Standard Process Model for Data Mining". In: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pp. 29–39. DOI: 10.1.1.198.5133.

Witkowski, Jens and David C. Parkes (2012). "A robust bayesian truth serum for small populations". In: *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pp. 1492–1498. ISBN: 9781577355687.

Wolfe, David T and Dana R Hermanson (2004). "The fraud diamond: Considering the four elements of fraud". In: *The CPA Journal* 74.12, p. 38. ISSN: 0732-8435.

Wolfers, Justin (2007). *Is there manipulation in the Hillary Clinton prediction market?* URL: http://www.overcomingbias.com/?s=manipulation+prediction+market (visited on 04/06/2017).

Wolfers, Justin and Andrew Leigh (2002). "Three Tools for Forecasting Federal Elections: Lessons from 2001". In: *Australian Journal of Political Science* 37.2, pp. 223–240. ISSN: 1036-1146. DOI: 10.1080/10361140220148115.

Wolfers, Justin and Eric Zitzewitz (2004). "Prediction Markets". In: *Journal of Economic Perspectives* 18.2, pp. 107–126. DOI: 10.3386/w10504.

— (2006a). *Five Open Questions About Prediction Markets*. Working Paper 12060. National Bureau of Economic Research. DOI: 10.3386/w12060.

— (2006b). *Interpreting Prediction Market Prices as Probabilities*. Working Paper 12200. National Bureau of Economic Research. DOI: 10.3386/w12200. URL: http://www.nber.org/papers/w12200.

Woodland, Linda M and Bill M Woodland (2011). "The Reverse Favorite Longshot Bias in the National Hockey League: Do Bettors Still Score on Longshots?" In: *Journal of Sports Economics* 12.1, pp. 106–117. DOI: 0.1177/1527002510368792.

Woudenberg, Fred (1991). "An evaluation of Delphi". In: *Technological Forecasting and Social Change* 40.2, pp. 131–150. ISSN: 0040-1625. DOI: 10.1016/0040-1625(91)90002-W.

Xintong, Guo et al. (2014). "Brief survey of crowdsourcing for data mining". In: *Expert Systems with Applications* 41.17, pp. 7987–7994. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2014.06.044.

Yang, JungHwan et al. (2016). "Why Are "Others" So Polarized? Perceived Political Polarization and Media Use in 10 Countries". In: *Journal of Computer-Mediated Communication* 21.5, pp. 349–367. ISSN: 1083-6101. DOI: 10.1111/jcc4.12166.

Yang, ShengYun, Ting Li, and Eric van Heck (2015). "Information transparency in prediction markets". In: *Decision Support Systems* 78.October, pp. 67–79. ISSN: 0167-9236. DOI: 10.1016/j.dss.2015.05.009.

Zhang, Yixiang et al. (2013). "Cognitive elaboration during wiki use in project teams: An empirical study". In: *Decision Support Systems* 55.3, pp. 792–801. ISSN: 01679236. DOI: 10.1016/j.dss.2013.03.004.

Zipfinger, Sabine (2007). *Computer-aided Delphi: an experimental Study of comparing round-based with real-time implementation of the method*. Trauner. ISBN: 3854992351.

Declaration of Authorship

I, Simon Andreas KLOKER, declare that this thesis titled, "Engineering Delphi-Markets for crowd-based Prediction: The FAZ.NET-Orakel and other cases" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what I have contributed myself.

Signed:

Date: