Charles University
Faculty of Social Sciences
Institute of Economic Studies

MASTER'S THESIS

Predicting Field Experiment Results in a Lab

Author: Bc. Kateřina Chadimová
Supervisor: PhDr. Lubomír Cingl, Ph.D.
Academic Year: 2016/2017
Declaration of Authorship

I hereby proclaim that I wrote my diploma thesis on my own under the leadership of my supervisor and that the references include all resources and literature I have used. I further declare that the thesis has not been used previously for obtaining any university degree.

I grant to Charles University a permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, May 19, 2017

__________________________
Signature
Acknowledgments

I would like to express my gratitude to PhDr. Lubomír Cingl, Ph.D. for supervising my work on this thesis, for his time, patience, and for his valuable advice throughout the year. My acknowledgement also deserves Dr. Jana Cahlíková for relevant comments and discussions. Finally, I would like to thank to my family for their unflagging support and love.
Bibliographical Record

Chadimová, K. (2017): Predicting Field Experiment Results in a Lab, Prague, 66 pp., Master’s Thesis (Mgr.), Charles University in Prague, Faculty of Social Sciences, Institute of Economic Studies, Supervisor: PhDr. Lubomír Cingl, Ph.D.

Bibliografický záznam

Chadimová, K. (2017): Predikce výsledků field experimentu v laboratoři, Praha, 66 s., Diplomová práce (Mgr.), Univerzita Karlova v Praze, Fakulta sociálních věd, Institut ekonomických studií, Vedoucí práce: PhDr. Lubomír Cingl, Ph.D.
Abstract

This thesis is aimed at forecasting of experimental results in a lab environment, investigating often discussed external validity of laboratory experiments. We run a novel laboratory experiment in which the subject pool is asked to make predictions on results of a certain field experiment. The collected data is analyzed using different accuracy measures, arriving at several interesting results. First, the forecast among the 94 subjects is quite informative about the actual treatment effects although its accuracy substantially varies based on a type of accuracy measure and a particular treatment. Second, the average forecast is either more accurate or at least comparable to the mean individual forecast, proving the presence of “wisdom-of-crowds” effect.

JEL Classification  C91, C92, C93, D03
Keywords  lab experiments, prediction accuracy, field experiments, behavioral economics

Range  113,409 characters

Author’s e-mail  kchadimova@gmail.com
Supervisor’s e-mail  lubomir.cingl@vse.cz
Abstrakt

Tato práce se zaměřuje na předpovídání experimentálních výsledků v laboratorním prostředí, čímž se zároveň věnuje často diskutované otázce externí validity laboratorních experimentů. Provedli jsme laboratorní experiment, ve kterém měli účastníci za úkol předpovídat výsledky určitého field experimentu. Sebraná data jsou analyzována pomocí několika různých nástrojů pro přesnost měření. Zjistili jsme, že předpovědi 94 účastníků experimentu poskytují dobrou informaci o reálných výsledcích field experimentu, ačkoliv jejich přesnost se mění v závislosti na použitém nástroji pro přesnost měření a určitém treatmenutu. Průměrné předpovědi jsou více přesné nebo alespoň srovnatelné s průměrnými individuálními předpovědmi, čímž je prokázána přítomnost efektu "moudrosti davu".

Klasifikace JEL

C91, C92, C93, D03

Klíčová slova

laboratorní experimenty, přesnost predikce,
field experimenty, behaviorální ekonomie

Rozsah práce

113 409 znaků

E-mail autora

kchadimova@gmail.com

E-mail vedoucího práce

lubomir.cingl@vse.cz
Contents

List of Tables ix

List of Figures x

Thesis Proposal xi

1 Introduction  1

2 Context  3
   2.1 Validity of Lab Experiments  3
      2.1.1 Internal vs External Validity  4
      2.1.2 Artificiality and the Type of Experiment  7
   2.2 Forecasts of Research Results  8
   2.3 Model  10
      2.3.1 Model’s Assumptions  10
      2.3.2 Further Specifications  11
      2.3.3 Derivation of the Variances  11
   2.4 Prediction markets  14
      2.4.1 Types of Prediction Market  14
      2.4.2 Applications and Evidence  16
      2.4.3 Accuracy of Prediction Markets  17

3 Framing in Field Experiments  19
   3.1 Framing effect  19
   3.2 Framing effect in Field Experiments  22
      3.2.1 Loss vs. Gain  22
      3.2.2 Moral Appeal  23
      3.2.3 Threat  24
      3.2.4 Social Norms  25
   3.3 The Behavioral Insights Team  27
List of Tables

2.1 Prediction Market: Contract Types .............................................. 15
5.1 Findings by Treatment: Response Rate ........................................ 40
5.2 Significances: Response Rate .................................................. 41
5.3 Rank Order: Response Rate ...................................................... 45
5.4 Accuracy of Forecasts versus Random Guesses: Response Rate . 46
5.5 Regression: Response Rate ..................................................... 49
5.6 Findings by Treatment: Registration Rate .................................. 50
5.7 Significances: Registration Rate ............................................... 51
5.8 Rank Order: Registration Rate .................................................. 54
5.9 Accuracy of Forecasts versus Random Guesses: Registration Rate 55
5.10 Regression: Registration Rate .................................................. 57
## List of Figures

2.1 Publication Trends ................................................................. 6

4.1 Field Experiment: Results ................................................... 35

5.1 Lab Experiment: Screenshot ............................................... 38
5.2 Power Calculations ............................................................. 42
5.3 Scatter Plot: Response Rate ................................................ 43
5.4 Box Plot: Response Rate ..................................................... 44
5.5 Individuals’ Accuracy: Response Rate ................................. 47
5.6 Individuals’ Accuracy in Treatments: Response Rate .......... 48
5.7 Scatter Plot: Registration Rate ............................................. 52
5.8 Box Plot: Registration Rate ................................................ 53
5.9 Individuals’ Accuracy: Registration Rate ............................. 56
5.10 Individuals’ Accuracy in Treatments: Registration Rate ....... 57

A.1 Baseline Letter (T1) ............................................................. II
A.2 Threat (T2) ................................................................. III
A.3 Social Info (T3) ............................................................. IV
A.4 Social Info + Threat (T4) ................................................ V
A.5 Moral Appeal (T5) ........................................................ VI
A.6 Moral Appeal + Threat (T6) .............................................. VII
Motivation Can the results of a field experiment be predicted in a laboratory experiment? With growing interest of general public in economic experiments it could be beneficial to forecast results of large field studies in a simple lab environment, obtaining first estimates of effectiveness of individual treatments and thus minimize potential unnecessary expenses. This thesis has therefore two related dimensions. First, it challenges an external validity of laboratory experiments and second, it follows and expands the literature on prediction accuracy in economics.

With a recent upsurge of field experiments the topic of external validity of laboratory experiments has been widely discussed. Indeed, laboratory experiments are often criticized for excessive artificiality that leads to results different from real world behavior (see e.g. discussion in Schram, (2005)). Therefore in this thesis we would like to scrutinize the extent to which predictions made in our lab experiment could be related to the actual effectiveness of field experiment treatments.

Economics has also a long tradition of studying accuracy of predictions. One example can be found in the literature on prediction accuracy in macroeconomics and finance (e.g. Ben-David et al. (2013)). Other example constitute prediction markets, i.e. markets where participants trade in contracts whose payoff depend on unknown future events. The information collected in these markets is often more accurate than the forecasts generated by more traditional methods such as opinion polls or expert judgment.

Newly, the attention has been drawn to forecasting of experimental results. DellaVigna & Pope (2016) analyze to which extent academic experts can anticipate the impact of different treatments and compare their forecasts to the sample of non-experts including students and online sample. Other examples can be found in Coffman & Niehaus (2014) who survey 7 experts on persuasion or in Sanders et al.
(2015) who ask 25 faculty and students from two universities to predict results of 15 experiments organized by the UK Nudge Unit.

**Hypotheses** Based on the paper by DellaVigna & Pope (2016) we state following research hypotheses:

Hypothesis #1: The average forecast predicts the experimental results quite well.

Hypothesis #2: There is some wisdom-of-crowds effect, i.e. the average forecast outperforms the individual ones.

Hypothesis #3: Accuracy defined as rank ordering treatments outperforms accuracy defined in absolute numbers.

**Methodology** We will run a controlled laboratory experiment with a student subject pool. We will analyze the data using standard forecast accuracy measures, i.e. mean absolute (squared) errors and their cumulative distribution functions.

**Expected Contribution** We would like to contribute to the growing body of literature focused on forecasting of experimental results. This is a relevant up-to-date topic since we can easily collect valuable information if it turns out that lab predictions dispose of a certain predictive value. Moreover, this thesis examines the extent to which laboratory results reflect field behavior, adding another important argument into the discussion of external validity. The fact that the topic of forecasting experimental results is rather new and unexplored also increases its relevance.

**Core bibliography**


Bott, K., Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2014). You’ve got mail: A randomised field experiment on tax evasion. NHH Norwegian School of Economics.


Chapter 1

Introduction

With a growing interest of general public in economic experiments it could be beneficial to forecast results of large field studies in a simple lab environment and obtain first rough but valuable estimates of effectiveness of individual treatments, thus avoiding unnecessary costs. Field experimentation with public and private sector is currently experiencing a rapid development as the cooperation between economists and organizations is relevant for both sides; careful application of insights from behavioral economics can significantly improve firm’s effectiveness or nudge hesitating people to act, whilst economists are given a unique opportunity to test various theories under the realistic setting. Nonetheless, large field experiments are typically complex to carry out as they require demanding preparation and the treatments might represent significant costs.

If it turns out that we can reliably forecast results of the field studies in a lab we will dispose of a powerful tool for generating first valuable estimates of variables of interest in a relatively short time compared to the commonly used pilot studies.

Therefore, the goal of this thesis is to test whether the results of a field experiment can be predicted in a simple laboratory experiment. In order to do that we design a novel lab experiment in which the recruited subjects (forecasters) are first introduced to the context and design of the specific field experiment and consequently they make predictions about its results. The topic is relevant from two reasons. First, as mentioned earlier, a proper and efficient design of field experiments presents an up-to-date challenge in the current ever-changing world in which time is precious and the experimenters are often under pressure to deliver the results in time, especially when cooperating with profit-oriented private entities. Second, by linking forecasts generated in a lab to a field be-
havior we address the widely discussed topic of external validity of laboratory experiments. Indeed, laboratory experiments are often criticized for excessive artificiality that might lead to unrealistic results. In this thesis we scrutinize the extent to which predictions made in the lab can predict the actual results of a particular field experiment.

The thesis is organized as follows. In the first part (Chapter 2) we define the context of our research question. The chapter begins with definition of experimental validity and explains the difference between internal and external validity. Then we summarize existing literature on forecasting research results and present a simple model on how forecasters form their beliefs about future research findings. We also mention the topic of prediction markets as it presents relevant and increasingly popular forecasting method that relates to our topic. The second part (Chapter 3) presents the power of framing effect and its application in various field experiments. The goal of this chapter is to highlight the importance of the manner in which the information is presented and introduce research that has been done so far in this area. In the third part (Chapter 4) the predicted field experiment by Fellner et al. (2013) is described in detail. We introduce the present reader to the institutional background underlying the research question, including experiment design with individual treatments and most importantly we depict the results of the field experiment. More specifically, Fellner with her colleagues vary the texting of letters sent to potential TV license fee evaders in Austria, collecting the data on response and registration rates invoked by the mailings. In the fourth part (Chapter 5) the design of our lab experiment is presented as we specify the task, pay-off function and subject pool. Furthermore, we analyze the collected data and scrutinize how closely it predicts the real results. Finally, the last part (Chapter 6) summarizes our main findings.
Chapter 2

Context

2.1 Validity of Lab Experiments

A lab experiment represents a suitable method of creating the proper counterfactual as it creates a control group directly through the means of randomization process. More importantly, it isolates the effect of interest from the influence of other factors that might be confounding, constituting a desired suitable environment for researchers. In order to be sure that the outcome of experiment is valid it is necessary to properly design its experiment’s structure. More specifically, in terms of validity of the experiment we have to distinguish two terms, i.e. internal and external validity.

A lab experiment constructs a small-scale environment in which an experimenter disposes of adequate control to ensure that treatment effects are measured appropriately. Such an environment consists of a set of agents \((1, \ldots, n)\) and commodities \((1, \ldots, k)\). Let’s suppose that each agent is endowed with a utility function \(u_i\), a technology (knowledge) endowment \(K_i\), and a commodity endowment \(w_i\). Therefore, each agent is defined by \(e_i(u_i, K_i, w_i)\) and the microeconomic environment is described by the collection of agents, \(\epsilon = (\epsilon_1, \ldots, \epsilon_n)\). Agents are assumed to possess consistent preferences and to make decisions in order to maximize their own well-being. To fully define the microeconomic environment we have to specify the institutional setting \(I\), including the appropriate message space \(M\), the allocation rules \(H\) and other characteristics that are relevant to the specific institution of interest. Therefore, we can now define the experimental system \(S = (\epsilon, I)\), consisting of the microeconomic environment and the institution. Specifically, agents choose messages and the institution determines allocations via governing rules (List 2007).
In order to measure reliably the behavioral principles between outcomes, institutions and preferences (i.e. to have a valid controlled microeconomic experiment) several sufficient conditions must be met.

**Assumption 2.1 (Non-satiation).** Given two identical options where they only differ in yield, the more profitable option will always be chosen.

**Assumption 2.2 (Saliency).** Subjects’ decisions are evaluated by a reward that is increasing for good outcomes and decreasing for bad outcomes.

**Assumption 2.3 (Dominance).** The reward structure dominates any subjective costs or values of participating in an experiment.

**Assumption 2.4 (Privacy).** Each subject is provided with its own payoff scheme.

**Assumption 2.5 (Parallelism).** Propositions about the behavior observed in a lab environment can be applied to non-laboratory microeconomies with similar ceteris paribus conditions as well (Smith 1982).

Once all the stated conditions are met, a controlled experiment in which the institution determines allocations and the agents follow consistently their preferences and choose messages is valid. Particularly, the first four percepts (i.e. non-satiation, saliency, dominance and privacy) are sufficient to allow for a controlled testing of economic theories ensuring the internal validity whilst the last assumption (i.e. parallelism) is enough to establish external validity.

### 2.1.1 Internal vs External Validity

**Definition 2.1 (Internal validity).** The internal validity of an experiment refers to the ability to draw confident causal conclusions from the research (Loewenstein 1999, p. 226).

In other words the internal validity assures that the observed differences on the dependent variable are caused directly by changing independent variable and not some other factor. Therefore, under the assumption of internal validity possible confounding factors (e.g. more than one independent variable is varying) should be minimized as much as possible. However, if an experimenter loses control over the study or if she uses unreliable measurements, the concept of the internal validity can be distorted (Bardsley 2010).

**Definition 2.2 (External validity).** External validity is “the extent to which the
results of a study can be generalized to and across populations, settings, and times.” (Johnson & Christensen 2000, p. 200)

The external validity is of especially high importance if we want to apply findings collected in a study outside of the experiment. Thus, a problem with the external validity of an experiment might arise whenever the recruited subjects behave in the different way than they would under the real world setting. This might be an example of behavioral bias that is also referred to as the “Hawthorne effect”. Indeed, the problem of external validity is often questioned by the critics of experiments in social sciences who claim that the artificiality omnipresent in a lab environment cannot lead to externally valid outcomes (Bardsley 2010).

Obviously there is tension between these two concepts as the internal validity yields results that are robust and replicable whilst external validity enables generalization of conclusions to situations that prompted the research. Card et al. (2011) classify all published lab and field experiments that were published in the top five economics journals from 1975 to 2010, finding that economic theory has played a more central role in the laboratory than in the field. More specifically, external validity might be significantly lowered as internal validity often forces the experimenters to abstraction and simplification to make the research more tractable. Thus the major threat to the external validity constitutes the artificiality in the lab setting whenever the laboratory incentives or institutions do not sufficiently correspond to those occurring in the real life situations they aim to study. The extent to which these parameters ought to be realistic depends on the type and goal of a particular experiment. Based on the long tradition of deductive reasoning and modeling in economics, the attention has been devoted mainly to the internal validity (Schram 2005). As Ortmann (2003) points out, the upsurge of field experimentation (Figure 2.1) has generated a vivid discussion about the external validity of laboratory experiments that is long overdue in economics.

1In 1920s, one of the early field experiment was run at the Hawthorne plant of the Western Electronic Company in Chicago. The experimenters tested whether brighter lighting at a workplace might increase workers' productivity. However, since the workers were aware about being part of the experiment their effort grew anyway independently on the illumination intensity. Later on, the data from the experiment was reanalyzed, suggesting the Hawthorne effect actually did not play a significant role in the results. Nevertheless, the Hawthorne Plant studies stand among the most influential social science research of the 20th century, as they have influenced subsequent research and methodology development to control for the confounding influence of scrutiny (Levitt & List 2011).
Figure 2.1: Number of Laboratory and Field Experiments Published in Top Five Economics Journals from 1975 to 2010

Source: Card et al. (2011).

Nevertheless, the debate on the artificiality in the economic experiments has been present in the literature for quite a long time. The criticism is often stemming from the presumption that the lab experiments do not reflect sufficiently the “real world” and thus do not provide relevant information about the actual economic behavior. Plott (1982) argues that experiments do not need to be realistic as long as they closely implement the tested theory. More specifically, Plott highlights the role of theory in simplification of the experimental process, claiming that the experimental design does not require to reflect the natural analog if the adequate accepted theory can be invoked. Following the critique in Cross (1980) of too large differences between experimental and outside-the-lab market, Vernon Smith argues that “empirical investigations of all aspects of parallelism between laboratory and field behavior are important. Similarly, our experimental and other investigations should not be confined to testing formal theory” Smith (1985, p. 288). Most importantly, this is the first time a leading experimental economist marks explicitly the external validity as a crucial topic. Loewenstein (1999) claims that a typical experiment possesses a high number of characteristics that are not relevant for decision making pro-
cess outside of the lab. Starmer (1999) does not consider the extent to which naturally occurring economic environment can be replicated in the lab to be “especially binding”, although he is aware of its limitations.

2.1.2 Artifi ciality and the Type of Experiment

In order to answer the question on how much the external validity is important for the experimentalists we have to take into account the goal of a particular study. Without that we are not able to properly evaluate the artifi ciality criticism.

Testing Theory

Experiments that are primarily designed to test various economic theories form the largest category. The main characteristics of this group is emphasis on their internal validity before the external validity and comprises two kinds of experiments. First, these are the studies used to establish institutions in the laboratory environment and test theories about their working. Second, these are experiments used to test various behavioral assumptions underlying a theory. Since the typical experiment aimed at testing theory is run under ceteris paribus conditions, we can expect the auxiliary hypotheses to fulfill these ceteris paribus conditions. To give an example consider an experiment in which we examine subject’s behavior in two treatments that differ from each other only in one specification. Under this setting it would not make any sense to presume that the subject understands the instructions differently in each treatment unless the opposite is explicitly stated. The same applies also to the experiments where identified causal relationship depends on a third variable that might be different outside of the lab.

Theory Stress Tests

Stress tests allow experimenters to test the domain of applicability of a theory while relaxing some of the assumptions in a controlled way. Conventionally, if a general theory is rejected in the lab there is no incentive to expect it would work well also in natural setting. What remains less clear is what it means when we find evidence that supports the theory since one can not simply deduce that such findings are directly applicable to the outside world. Therefore, it is important to move further in the external validity testing and use various
instruments, including computer simulations, empirical analysis of real world data and field experiments.

Searching for empirical regularities
It often happens that a proposed theory is rejected in the laboratory or that no economic theory is available at all for a proposed research topic. In this case, data collected in experiments under carefully controlled circumstances might be used to establish stylized facts which could be formulated either in a form of observed phenomena or observed causal effects. Nevertheless, it is crucial to pay some attention to the problem of external validity as in the contrast with testing theories. The data must be relevant for the situation an experimenter is interested in since she must find a proper guidance in the real world that she aims to study.

Advising policy makers
Recently more and more policy makers has contacted experimental economists for advice. Although it is not obvious at first sight, a lot of these “policy advising” experiments are guided by theory and therefore the emphasis on their external validity is not great. This is particularly true for the case of auction experiments. Nevertheless, many research questions of this kind cannot be simply answered by theory and therefore we should avoid artificiality as much as possible. However it might be difficult to assess whether the level of external validity is high enough in individual studies. Perhaps, as the policymakers find such studies useful, we can take external demand for laboratory experiments as a positive measure of their external validity (Schram 2005).

2.2 Forecasts of Research Results
In this section we summarize a relatively small but growing literature on forecasts of economic research results.

Erev et al. (2010) that organize three choice prediction competitions is one of the first example of predicting experimental results. The decision experiments constitute of three choice tasks: i) one shot decisions from description, ii) one shot decisions from experience and iii) repeated decisions from experience. Having collected the data from the experiments, the organizers posted them online complemented by several baseline models and asked other labo-
ratory experimenters to predict the results of the second experimental session. Thus the predictions were made using learning models trained on data.

In contrast Erev *et al.* (2010), DellaVigna & Pope (2016b) are more interested in whether academic experts are able to make quick and intuitive forecasts such as are often done in an informal consulting, advising, or mentoring sessions. They run a simple prediction market as a supplement study of their large-scale field experiment in which they compare the effect of three categories of motivators: i) standard incentives, ii) behavioral factors like e.g. social preferences or present bias and iii) non-monetary inducements. The results of the field experiment are then compared to forecasts made by 208 academic experts, concluding that the average forecast of experts predicts the experimental results quite well. They also find a strong wisdom-of-crowds effect as the average forecast outperforms 96% of individual forecast. In the follow-up study DellaVigna & Pope (2016a) compare predictions made by experts to predictions made by PhD students and non-experts (i.e. undergraduates, MBAs and an online sample). Not surprisingly, experts perform better as a group than non-experts. However if the accuracy of prediction is defined as rank ordering treatments then there are no significant differences in predictions between both groups.

Another example can be found in Coffman & Niehaus (2014) who contact 8 researches specialized in persuasion from across economics and psychology and ask them to make predictions of buyers’ evaluation. Similarly, Sanders *et al.* (2015) survey 25 faculty and students from two universities on 15 selected experiments that are organized by the UK Behavioral Insight Team (Nudge Unit)\(^2\). Finally, Groh *et al.* (2016) elicit expectations of academics and development professionals on a randomized experiment run in Jordan in which female college graduates were randomly matched with a soft skills training program.

As suggested in DellaVigna & Pope (2016b), the topic of predicting research findings calls for a more systematic work as it provides an interesting research opportunity. Moreover, the information collected in these prediction markets might be useful in improving decision making both in business and public policy.

\(^2\)For more information about BIT see Section 3.3.
2.3 Model

Within this section we present simple model developed by DellaVigna & Pope (2016a) on how forecasters form their beliefs about future research findings. To our knowledge this is the first model of its kind and certainly more work should be done in order to fully understand the underlying mechanism behind making predictions. Overall, the model is able to reproduce several patterns that could be found in the forecasted data, including individual accuracy versus the wisdom-of-crowds, differences between absolute error and rank-order correlation or performance of the so called “superforecasters”. Furthermore, it is also able to capture differences between particular groups defined e.g. according to the level of their expertise, i.e. how much they are familiar with the research topic (DellaVigna & Pope 2016a).

2.3.1 Model’s Assumptions

Let us suppose there is an agent $i$, making forecasts about the results in experimental treatments $k = 1, \ldots, K$. Let $\theta = (\theta_1, \ldots, \theta_K)$ be the outcome in the $K$ treatments that is unknown to the agent. Motivated by the incentives in the survey (in our specific case the participants are motivated by a financial reward), the agent aims to minimize the squared distance between the forecast $f^i_k$ and the result $\theta_k$. Agents start with a non-informative prior and consequently agent $i$, with $i = 1, \ldots, I$, draws a signal $s^i_k$ about the outcome of $k$ treatment, i.e.

\[
s^i_k = \theta_k + \eta_k + v_i + \sigma_i \epsilon^i_k. \tag{2.1}
\]

Therefore, the difference between the signal $s^i_k$ and the truth $\theta_k$ is affected by three factors, each i.i.d. and independent from the other components, where

i) $\eta_k \sim N(0, \sigma^2_\eta)$ is a deviation for treatment $k$ that is common to all forecasters,

ii) $v_i \sim N(\mu, \sigma^2_v)$ is a deviation for forecaster $i$ that is common across all treatments, yielding a possible bias term if $\mu \neq 0$,

iii) $\sigma_i \epsilon^i_k$, with $\epsilon^i_k \sim N(0, 1)$, is the idiosyncratic noise component, with heterogeneous $\sigma_i$ that is lower for more accurate forecasters.
Moreover, we assume that $\sigma_i^2$ is independent of $\epsilon_k^i$ and that the idiosyncratic variance $\sigma_i^2$ follows an inverse gamma distribution, i.e. $\sigma_i^2 \sim IG(\alpha, \beta)$.

Next, we assume the agent to be unaware of the systemic bias $\mu$. If we combine this assumption with the uninformative prior then the signal $s_k^i$ represents an agent’s best estimate, i.e. $f_k^i = s_k^i$. In other words, the agent’s best estimate minimizes the subjective expected squared distance between the forecast and the result in the treatment $k$.

### 2.3.2 Further Specifications

Idiosyncratic noise presented in the forecasts is captured by the error term $\epsilon_k^i$. Furthermore, the forecasters differ in the extent of idiosyncratic noise with some forecasters who provide less noisy forecasts (lower $\sigma_i$) when compared to others. This heterogeneity affects the correlation of errors across treatments. More specifically, if $\sigma_i$ does not significantly differ across forecasts, the individual absolute error in one treatment will have little predictability for the absolute error of the same person in another treatment since the forecast error arises from noise that is similar across all forecasters. Contrary, in case that there are some forecasters with significantly lower $\sigma_i$ in comparison with other forecasters there will be cross-treatment predictability. In other words, the forecasters who perform well in one treatment are likely to have low $\sigma_i$ and therefore they are also likely to do well in another treatment. To summarize, heterogeneity in $\sigma_i$ can indeed capture the results on revealed forecasting ability.

We include the additional error terms $\eta_k$ and $v_i$ into the model from the following reasons. First, some treatments might be harder to forecast than others and potential cross-treatments differences arising from this reason cannot be explained by the idiosyncratic error. Therefore, the term $\eta_k$ allows for such differences, potentially accounting for an incorrect common reading of the literature or context for a particular treatment or an unusual experimental finding. Second, there might be another reason why the predictions differ more than it could be explained by idiosyncratic noise or difficulty to predict individual treatments. Thus, the term $v_i$ captures agent $i$ being more optimistic or pessimistic about the results of all treatments.

### 2.3.3 Derivation of the Variances

In order to derive variances that might be used for eventual calibration of the model we proceed like this. First, we compute variance of the idiosyncratic
2. Context

noise component

\[ \text{Var}(\sigma_i e_k^i) = E(\sigma_i^2 (e_k^i)^2) - [E(\sigma_i^2 e_k^i)]^2 = E(\sigma_i^2)E((\epsilon_k^i)^2) - E(\sigma_i^2)^2 E(\epsilon_k^i)^2 = E(\sigma_i^2) \cdot 1 - E(\sigma_i^2)^2 \cdot 0 = E(\sigma_i^2) \]

Second, the cross-sectional variance equals

\[ \text{Var}(s_k^i - \theta_k) = \text{Var}(\eta_k) + \text{Var}(v_i) + \text{Var}(\sigma_i e_k^i) = \sigma_n^2 + \sigma_v^2 + E(\sigma_i^2) \]

Third, the wisdom-of-crowds variance satisfies

\[
\begin{align*}
\text{Var}(s_k^i - \theta_k) &= \frac{1}{I^2} \left[ \sum_{i=1}^{I} \text{Var}(\eta_k + v_i + \sigma_i e_k^i) + 2 \sum_{i<j} \text{Cov}(\eta_k + v_i + \sigma_i e_k^i, \eta_k + v_j + \sigma_j e_k^j) \right] \\
&= \frac{1}{I^2} \left[ I(\sigma_n^2 + \sigma_v^2 + E(\sigma_i^2)) + 2 \sum_{i<j} \text{Var}(\eta_k + v_j + \sigma_j e_k^j) + \text{Cov}(\eta_k, v_j + \sigma_j e_k^j) \right] \\
&\quad + \text{Cov}(\eta_k, v_i + \sigma_i e_k^i) + \text{Cov}(v_i + \sigma_i e_k^i, v_j + \sigma_j e_k^j) \\
&= \frac{1}{I^2} \left[ I(\sigma_n^2 + \sigma_v^2 + E(\sigma_i^2)) + I(I-1)\sigma_n^2 \right] \\
&= \sigma_n^2 + \frac{1}{I}(\sigma_v^2 + E(\sigma_i^2))
\end{align*}
\]

Furthermore, for the average-bias variance we obtain

\[
\begin{align*}
\text{Var}(s_i^i - \theta) &= \frac{1}{K^2} \left[ \sum_{k=1}^{K} \text{Var}(\eta_k + v_i + \sigma_i e_k^i) + 2 \sum_{k<l} \text{Cov}(\eta_k + v_i + \sigma_i e_k^i, \eta_l + v_i + \sigma_l e_l^i) \right] \\
&= \frac{1}{K^2} \left[ K(\sigma_n^2 + \sigma_v^2 + E(\sigma_i^2)) + K(K-1)\sigma_v^2 \right] \\
&= \sigma_v^2 + \frac{1}{K}(\sigma_n^2 + E(\sigma_i^2))
\end{align*}
\]
Having calculated these expressions, we can proceed further to solve for $E(\sigma_i^2), \hat{\sigma}_v^2$ and $\hat{\sigma}_\eta^2$

\[
\hat{\sigma}_\eta^2 = \frac{1}{I-1}(IVar(s_k - \theta_k) - Var(s_k^i - \theta_k))
\]

\[
\hat{\sigma}_v^2 = \frac{1}{K-1}(KVar(s_i - \theta) - Var(s_k^i - \theta_k))
\]

\[
E(\sigma_i^2) = Var(s_k^i - \theta_k) - \hat{\sigma}_\eta^2 - \hat{\sigma}_v^2
\]

Since we assume $\sigma_i^2$ to have inverse gamma distribution $IG(\alpha, \beta)$, using the standard properties we arrive to

\[
\alpha = 2 + \frac{E[\sigma_i^2]}{Var[\sigma_i^2]}
\]

\[
\beta = E[\sigma_i^2] \left(1 + \frac{E[\sigma_i^2]}{Var[\sigma_i^2]}\right)
\]

Finally, if we compute $E[\sigma_i^2]$ with the use of calibration method, we can vary $Var[\sigma_i^2]$ through the implied values of $\alpha$ and $\beta$ to match the correlation of absolute error across the given treatments.
2. Context

2.4 Prediction markets

Definition 2.3 (Prediction Market). “Prediction market is a market where participants trade in contracts whose payoff depends on unknown future events” (Wolfers & Zitzewitz 2004, 109).

Prediction market (also known as “information market” or “event futures”) is a relatively new and emerging form of financial market that is created for the purpose of trading the outcome of events. Importantly, they can be used for aggregation and revelation of disperse information into efficient forecasts of uncertain future events. According to the efficient markets hypothesis the market prices reflect what the traders think the probability of the event is, i.e. the market prices are the best predictor of the event. Equivalently no other combination of available polls or information can be used to improve on these market-generated forecasts. At the same time the efficient market hypothesis does not assume all the individuals in a market to be rational, it only suffices that the marginal trade in the market is led by rational traders. Although it is unlikely that prediction markets are literally efficient, there is a convincing number of successes in these markets (both within firms and with regard to public events like e.g. presidential election) that have awaken curiosity among economists, constituting a promising field to be studied (Wolfers & Zitzewitz 2004).

2.4.1 Types of Prediction Market

As has been stated above, on of the main characteristic of the prediction markets are payoffs that are tied to unknown future events. Naturally, the manner in which the payoff is linked to the future event matters as it might substantially affect a range of parameters of the market’s expectations (Wolfers & Zitzewitz 2004). Throughout this section we will refer to the market as a representative agent with a set of expectations. Nonetheless, the reader should be aware that there are subtle yet important differences between e.g. the market’s median expectation and the median expectation of market participants.

The following table summarizes the three main types of contracts that might be traded in prediction markets. These contracts provide information about the market’s expectation of a specific parameter, i.e. a probability, mean and median, respectively. Nevertheless, prediction markets might be used to assess the level of uncertainty about these expectations (Wolfers & Zitzewitz 2004).
Let us assume that we have a winner-takes-all type of contract. Under this setting, the contract costs a certain amount $p$ and yields a certain payoff, e.g. $1$, if and only if a specific event occurs, e.g. a candidate wins an election. Assuming risk neutrality, the price $p$ on a such type of prediction market represents the market’s expectation of the probability that the particular event will occur. In other words, the price of a winner-takes-all security represents a state price that is equal to an estimate of the future event’s probability. We can use the argument of risk neutrality here as the stakes in prediction markets are usually small enough to enable us to assume that investors are not averse to the idiosyncratic risk involved. Nevertheless, it might happen that the future event is correlated with investors’ marginal utility of wealth. In this case, the probabilities and state prices might differ (Wolfers & Zitzewitz 2004).

As opposed to the previous “all or nothing” type, an index contract yields payoff that is continuously varied based on a factor of interest, e.g. the percentage of the vote received by a particular candidate. Obviously, the price of such a contract equals to the mean value assigned to the outcome by the market itself (Wolfers & Zitzewitz 2004).

<table>
<thead>
<tr>
<th>Contract</th>
<th>Example</th>
<th>Details</th>
<th>Reveals market expectation of ...</th>
</tr>
</thead>
</table>
| Winner-takes-all | Event y: Al Gore wins the popular vote | Contract costs $p
Pays $1 if and only if event $y$ occurs
Bid according to value of $p$ | Probability that event $y$ occurs, $p(y)$ |
| Index          | Contract pays $1 for every percentage point of the popular vote won by Al Gore | Contract pays $y$. | Mean value of outcome $y$: $E[y]$ |
| Spread         | Contract pays even money if Gore wins more than $y^*$% of the popular vote. | Contract costs $1
Pays $2 if $y > y^*$
Pays $0 otherwise.
Bid according to the value of $y^*$. | Median value of $y$. |

A spread contract is characterized by traders bidding on the specific threshold, determining whether an event occurs or not. Continuing with the Al Gore example from Table 2.1, the traded contract will be a “spread” one when a candidate receives more than a certain percentage of the popular vote, exceeding the stated cutoff. Another example might be a point-spread betting, where the contract is either that a football team will win by at least a specific number of points or will not.

The spread contract pays fixed amount of money if the future event will occur but the size of the spread can vary. Furthermore, when this type of contract is combined with an even-money bet, i.e. a prize is doubled for winners whilst losers receive zero, the contract reveals market expectation of the median outcome, constituting the only fair bet if a payoff is as likely to occur as not (Wolfers & Zitzewitz 2004).

2.4.2 Applications and Evidence

Iowa Electronic Market, run by University of Iowa is one of the most famous prediction markets at least among the economists. In 1988, the first Iowa experiment on the presidential election won by Bush or Dukakis was launched, trading in an index contract that would pay 2.5 cents for each percentage point of the popular vote. Later on, it has released markets on the 2003 California gubernatorial election, the 2004 US presidential election or the 2004 Democratic presidential nomination. Inspired by the University of Iowa project, other universities from all over the world have opened their own prediction markets on the local elections like e.g. Austrian Electronic Market or the University of British Columbia Election Stock Market.

Another category consists of web-based prediction markets that are often organized by companies that specialize in trading and gambling services. Interestingly, there exist also pseudo-markets where one might trade in virtual currency. For example The Hollywood Stock Exchange runs such a market in which traders speculate on movie-related events such as e.g. opening-weekend performance or awarding Oscars.

To give another example, Goldman Sachs and Deutsche Bank have recently opened their own markets on the likely outcome of future readings of economics statistics such as retail sales, business confidence or industrial production. There are also event markets that are focused on forecasting private-sector
returns or used by private firms as a business forecasting tool (Wolfers & Zitzewitz 2004).

2.4.3 Accuracy of Prediction Markets

Without any doubts the most appealing issue concerning prediction markets is their efficiency as predictive tools. Theoretically two following assumptions are essential to a working event market:

i) The number of traders must be large enough so that the available information could successfully aggregate and correctly predict the outcome of the event of interest.

ii) The underlying market mechanism must facilitate a disperse information aggregation in order to generate the prevailing market price that is reliable statistic for the traders’ collective information.

The question whether real markets are able to fulfill these two assumptions is the issue of behavioral origin and can be best answered by studying both individual trader activity and market dynamics (Berg et al. 2008).

Evidence

Berg et al. (2008) compare data collected on the Iowa Electronic Markets with poll data from the subset of national elections. Comparing average absolute errors of the estimates, the authors reveal that the market outperformed polls in 9 of 15 elections with the average poll and market error being equal to 1.91% and 1.55% respectively. Moreover, in a few cases the market performed significantly better, concluding that the market itself is able to yield very accurate predictions.

In their study, Servan-Schreiber et al. (2004) ask a question about how much extra accuracy in prediction market could be gained from using real instead of extra money. They use online domains TradeSports.com (real-money) and NewsFutures.com (play-money) to create a real-world experiment incorporating traders interested in National Football League (NFL) betting. Unsurprisingly, they find that both types of markets provide useful estimates of average beliefs about the probability of future events. More interestingly, they conclude that the play-money prediction market perform as well as the real-money market. The authors suggest this could be due to two antagonistic factors as i) the real-money market might better support discovery of relevant information whilst ii)
the play-money market might support more efficient aggregation of information. Other examples could be found in Plott & Chen (2002), Ait-Sahalia et al. (2001) or in Plott & Sunder (1988).

The papers listed above provide a field evidence across several domains that prediction markets can indeed serve as accurate predictor of probabilities. More importantly, the following literature suggests that this evidence could be directly linked to the theory.

Wolfers & Zitzewitz (2006) explore several prediction market models, proving that the prices on these markets aggregate beliefs very well. Nevertheless, prediction market prices aggregate information into useful forecasts only if the traders are well-informed, having access to the relevant information. Moreover, the efficacy of these forecasts can be lowered for prices close to $0 or $1. This situation might occur when the distribution of beliefs is either exceptionally disperse, or when volumes traded are affected by an unusual degree of risk-acceptance or constrained to some extent. Nonetheless, they conclude that although prediction markets prices could be sometimes biased, they still exhibit significant prediction powers.

The study was inspired by Manski (2006) who criticizes the lack of existing theory supporting the interpretation of prediction market prices of binary options as predictions of the probability of future events. Eventually, Wolfers & Zitzewitz (2006) show that Manski’s specific example represents actually a worst-case scenario of prediction market model.

Similarly, Ottaviani & Sorenson (2007) analyze a model of simple market for a binary event in which traders take positions in two Arrow-Debreu contingent assets with payoff equal to one dollar if the outcome of interest is realized. Since the traders are not fully experienced with such situations, they are assumed to have heterogeneous beliefs having origin in two sources. First, they have subjective prior beliefs that are uncorrelated with the realization of the outcome. Second, they dispose of various information. Since the information is correlated with the outcome it has an objective nature. The authors show that the model’s rational expectations equilibrium price underreacts to information.
Chapter 3

Framing in Field Experiments

3.1 Framing effect

Definition 3.1 (Framing effect). “Framing effect is said to occur when equivalent descriptions of a decision problem lead to systematically different decision” (Sher & McKenzie 2006, p. 468).

Generally, economics and other social sciences are based on the assumption of human rationality. Although there is still vivid debate about the precise meaning of rationality, it is widely agreed that rational choices should follow some basic requirements such as consistency and coherence. However, there exist decision problems in which people systematically depart from rationality axioms, i.e. consistency and coherence, and these violations can be linked with the psychological patterns supervising the perception of decision problems and the evaluation of options.

A decision problem is characterized by first a set of possible options or acts from which one must choose, second by a set of possible outcomes or consequences arising from choosing an option and third by the contingencies or conditional probabilities that link outcomes to acts. Therefore, a “decision frame” refers to the decision maker’s notion of the acts, outcomes and conditional probabilities that are altogether aligned with a particular choice. Moreover, the way in which one adopts the decision frame is not only affected by the problem formulation but also by the norms, habits, experience and personal characteristics of the decision maker.

Theory of rational choice suggests that option ranking should not reverse with changes of frame. In the real world however, human ability to perceive and decide is limited and thus imperfect and therefore systematic reversals
of preference following variations in the framing of acts, contingencies or outcomes, might arise, being at odds with the rationality assumption. In other words, preferences are indeed dependent on the formulation of decision problems (Tversky & Kahneman 1981).

Let us illustrate the problem of reverse preferences due to the framing on the example mentioned by Tversky and Kahneman:

Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows. Which of the two programs would you favor?

**Problem 1:**
If Program A is adopted, 200 people will be saved. [72%]
If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved. [28%]

**Problem 2:**
If Program C is adopted 400 people will die. [22%]
If Program D is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. [78%]

Although the problems are essentially the same, there is a striking difference in their formulation. In Problem 1, the outcomes are defined in terms of lives saved whereas in the Problem 2 in terms of lives lost. Thus, in the Problem 1 the majority of people is risk averse preferring Program A over B as they consider secure saving of 200 human lives more feasible to relate than the risky prospect of equal expected value, i.e. one-in-three chance of saving 600 lives. However, in the Problem 2, the majority of people is risk taking when favoring two-in-three probability that 600 people will die over the certain death of 400 people. Thus, the simple framing of the problem in terms of gains (positive) and losses (negative) generates risk averse and risk taking choices respectively, leading to inconsistency responses to the effectively identical problems (Tversky & Kahneman 1981).

Indeed economists together with psychologists agree that the manner in which a decision problem is presented matters. E.g. Thaler (1980) defines the endowment effect as the underweighting of opportunity costs. In other
words, a loss is generated if a good is removed from the endowment whilst a gain is experienced if the same good is added to the endowment without it. Therefore, an economic agent appreciates more goods that are included in his or her endowment compared to those excluded. Thaler illustrates this by an example in which people were willing to accept a 0.001 risk of sudden death in exchange for a minimal compensation that was by one or two orders of magnitude higher compared to the amount people approved to pay in order to eliminate an identical existing risk. Likewise, Kahneman et al. (1990) run an experiment in which they construct a market for coffee mugs. Despite the Coase theorem predictions, the observed volume of mugs traded is significantly less even when subjects dispose of opportunities to learn. Thus, as in contrast with the theory, willingness to accept the risk is markedly higher than willingness to pay.

Another example might be status quo bias. Samuelson & Zeckhauser (1988) designed a series of decision-making experiments, finding that subjects tend to choose status quo option more frequently than could be expected from the canonical model. The participants faced the questionnaire based on decision problems, each requiring a choice from among a fixed number of alternatives. Under the neutral framing, subjects were asked to choose from a set of potential equal alternatives whilst under the status quo manipulation, one of the options was marked as a status quo and the others were presented as its alternatives, while controlling for preference and keeping constant the set of choice alternatives. However, the level of status quo bias was affected by the individual’s observable preference and by the number of alternatives in the choice set.

Although there is a relatively rich lab evidence on the power of the framing effect, the results are often questioned by many economists claiming that the lab setting cannot generate real values stakes, subjects or environment, causing the experimental findings to be unrepresentative. One possible way how to address this critique might be field experimentation. Field disposes of advantage of that e.g. both status quo and entitlements to endowments arise naturally there and thus the treatment effect creates more authentic responds. Nonetheless, complexity and opacity of field environment stemming from the high number of present factors might make the process difficult for the researches. Therefore, it is crucial to follow the methodology carefully to be able to shed some light on the economic effects of framing manipulations in the naturally-occurring markets.
3. Framing in Field Experiments

3.2 Framing effect in Field Experiments

In the following section we present several field experiments that employ framing manipulations in order to illustrate the power of manner in which a decision is presented. Indeed, the results of a field study might significantly differ based on the design of particular treatments.

3.2.1 Loss vs. Gain

Hossain & List (2012) run a natural field experiment in cooperation with a high tech Chinese enterprise, producing and distributing consumer electronics. The research question was whether different bonus schemes might substantially influence workers’ productivity. In order to address the issue, they created two treatments, i.e. positive (“reward”) and negative (“punishment”). Under the “reward” scheme, workers were informed they will be paid a bonus at the end of the pay period if the week’s average per-hour production exceeds a certain level. On the other hand, under the “punishment” scenario, employees were given a provisional bonus in advance, facing the possibility of its retraction if they were not able to reach the productivity benchmark. Generally, both the bonus schemes had positive effect on the workers’ productivity. However, the “reward” scheme invoked larger effect by 1% compared to the “reward” scheme. In other words, the total size of the framing effect, i.e. a slightly different formulation of the decision problem caused total productivity to increase by 1%.

Clearly, this is a great example of a presence of loss aversion in a naturally occurring labor market. Moreover, the authors found the effect to persist over the entire sample period of six months, suggesting economically significant long-term growth.

The results are important from several reasons. First, they support the importance of framing of incentives and show how the findings behavioral economics can be applied in the workplace with impressive conclusions. Second, “results speak to the literature in the broader social sciences on how social structure and institutions serve as important constraints influencing behavior” (Hossain & List 2012, p. 2153)

Therefore, the experiment by Hossain and List represents a field counterpart to lab experiments by Tversky, Kahneman or Thaler, that have highlighted the difference in gains/losses framing for many years. Indeed, the field evidence
might now mitigate the critiques who discard the lab findings because of omnipresent artificiality.

### 3.2.2 Moral Appeal

Bott *et al.* (2014) address the problem of tax evasion in the experiment that was conducted in Norway. As Norwegian tax payers are expected to self-report their foreign income, they might be tempted to reduce the final sum reported. This is especially true in a situation of ineffective control and malfunctioning enforcement systems. Although obviously the total amount of misreported taxes could be reduced by rising the detection probability and tightening of potential sanctions, implementation of appropriate measures might be very costly. Therefore, it may be beneficial to get insight into other possible ways how to combat tax evasion that would be less expensive and more straightforward to implement.

In the experiment itself, tax payers received an information letter posted by the tax administration right before they were obliged to submit their tax return for the fiscal year 2012. The control group received a base letter, providing the information about the purpose of reporting foreign income and wealth. The treatment lied in adding a moral appeal into the base letter, either in the form of reminder that the most Norwegian tax subjects report their income correctly and completely or in the form of statement that collected taxes are used to finance crucial public goods and services such as education or research. In order to compare the power of moral appeal with the threat of being detected, the experimenters designed another letter which was targeted at increasing the taxpayers’ subjective probability of detection.

The subjects involved in the experiments were those who were the most likely to under report their foreign income based on the evidence from the previous fiscal year. The careful randomization of subjects into the treatment and control groups assured that any systematic variation in the self-reported foreign income must been the result of the different framing. In fact, the authors found that the treatment letter almost doubled the reported average foreign income compared to the base letter. Interestingly, although moral appeal and increase in perceived probability of detection generated approximately the same effect, they performed in different manner. Whilst moral appeal framing did not increase the number of tax subjects reporting their foreign income, it worked on the intensive margin, i.e. it caused an increase in the amount reported of
those people who report any foreign income. On the other hand, the higher probability of detection worked mainly on the extensive margin, inflating the number of tax payers who report any foreign income.

The results are thus interesting from various reasons. First, although there is a high number of lab studies concerning moral incentives in making economic decisions (see e.g. Andreoni & Miller (2002)) the findings strengthen the role of moral motivation in everyday life. More interestingly, the authors show that morality matters even in a high stake environment in which subjects can easily generate high profits by deceitful behavior. Second, the intervention in a form of adding few sentences into the base letter is essentially with no expenses but it can have tremendous effect on the collected taxes. In the Norwegian experiment the framing itself increased the self-reported income by approximately 150 million NOK. This is an interesting result supporting the importance of behavior economics.

3.2.3 Threat

Similarly, in their paper Hallsworth et al. (2015) explore the performance of framing effect in the high stakes environment. Their assumption is that a lot of people engage in dishonorable behavior simply because they prioritize failing to act before making an active choice. This might be the case when paying taxes, fines or repaying debts. Thus, the experimenters cooperated with the Government Tax Office (Her Majesty’s Revenue and Customs) which is responsible for the tax and debt collection in the United Kingdom and focused on the disbursement of government benefits. Generally, there exist many benefit programs and typically a number of certain conditions is required to be fulfilled in order to be eligible for receiving benefits. Likewise, recipients are expected to notify the disburser whenever the conditions change, losing benefits entitlement. Inevitably, this setting creates a space for making a profit by choosing an omission strategy anytime the recipient fails to report the change - either consciously or unconsciously.

Particularly, Hallsworth et al. (2015) focused on collection overpaid Tax Credits, consisting mainly of Child Tax Credit and Working Tax Credit and they sent a letter to the people who owed money to the government. The control group received an ordinary letter, containing basic information about the size of debt and how to repay whilst the treatment letter included additional various short messages. These extra messages took essentially two forms; first,
3. Framing in Field Experiments

Omission to Commission interventions and second, offering of help and lowering subjective barriers to resolve the recipient’s debt situation. Under the first scheme, the additional message was as follows: “Previously, we treated your lack of response as an oversight. Now, if you do not call [telephone number], we will treat this as an active choice.”

Under the second scheme, the letter provided debtors with more detailed information about call center opening times (“Opportunity” framing), expressing authority interest in resolving the issue (“Reciprocity” manipulation), proposal to make a plan to call the tax authority (“Planning” manipulation) and summarizing all the main points in the letter (“Summary Box” manipulation). The experimenters focused on period of 30 days of being sent the letter as after the month debtors might receive another letter asking for payment. Approximately 12% of people in the control group repaid the debt after receiving the base letter. Interestingly, the “Reciprocity”, “Planning” and “Summary Box” framing did not have any substantial effect on repayment rates whilst “Opportunity” treatment persuaded 2.2% more people to make the payment compared to the control subjects. On the other hand, “Omission to Commission” scheme invoked approximately an 11 percentage point increase in payment rate, constituting a 92% treatment effect and generating over $1.3 million of new yield. This is a relatively high effect, affecting large amounts of money. Therefore, the authors suggest that policymakers should be careful when designing policies or services and they should minimize opportunities for exploitation of omission options. As they put: “In other words, if given the chance, people will select an omission option because they think they will be judged less harshly by others as a result of acting dishonestly” (Hallsworth et al. 2015, p. 11).

3.2.4 Social Norms

In order to mitigate disturbing consequences of changing climate, many people regard improved energy efficiency as an effective tool how to substantially reduce energy demand and decrease greenhouse gas emissions. Over many years economists have favored relative prices, subsidies or allowances as the main instruments of correction of negative market externalities. However, as it turns out, non-price interventions “nudge” consumers to save energy efficiently and to think about their consumption, having effects on consumer behavior that might be similar to the large changes in relative prices. The main advantage of those non-price incentives is the minimal implementation cost. Yet, it is
crucial to design them appropriately in order to be able to generate interventions that are scalable and can significantly affect a representative population Allcott (2011).

Allcott (2011) is especially interested in how social norms can affect individual behavior. He mentions three different ways how the information on neighbors’ electricity consumption might reduce individuals’ demand. First, some individuals might enjoy being presented as more frugal compared to their neighbors. Second, since externalities from power plant greenhouse gas are not reflected in electricity prices, a lot of consumers might realize that energy conservation helps in providing a public good. Last but not least, such information might significantly facilitate social learning.

In order to test the hypotheses, Allcott evaluated the OPOWER pilot program operating in the USA. The participated households were randomly assigned to the control and treatment groups of the similar size. The treatment group received regular Home Energy Reports, a several-page letters consisting of Social Comparison Module and Action Steps Module. The Social Comparison Module compares the household’s electricity consumption over the past year to the mean of its comparison group and the 20th percentile whilst Action Steps Module suggests possible practical steps that might be implemented by households in order to reduce the energy consumption.

The studied period lasted for about one and half year, when 60% of households in the treatment group were assigned to receive the letters monthly whereas 40% of them received them quarterly. The average treatment effect was estimated to be between 1.9-2.0 percent below baseline. However, in the group that received the Reports quarterly this effect decayed to some extent in the months between Reports as the information decays seasonally or because the attention of the consumers weaken over time. Unsurprisingly, the effect of the intervention was strongest for households that exhibited highest baseline consumption. This finding is consistent with the phenomenon of “boomerang effect”, suggesting that households with low baseline consumption are more inert to learn the social norms or they might even increase their consumption. Thus, simple framing in terms of social norms and feedback provision (non-price incentives implementation) improved the program’s welfare implications and generated effects that are comparable to the traditional large monetary subsidies for energy efficient durable goods. Importantly, Allcott proves that social norms can be source of fundamental changes in consumer behavior even at a population scale.
3.3 The Behavioral Insights Team

The Behavioral Insights Team - also known informally as the Nudge Unit - is a social purpose company based in the United Kingdom and partly owned by the UK government (Cabinet Office), Nesta and its own employees. It is the first governmental institution in the world which purpose is the practical application of academic behavioral research in order to improve public services. Particularly, its stated objectives are:

- making public services more cost-effective and easier for citizens to use
- improving outcomes by introducing a more realistic model of human behavior to policy
- wherever possible, enabling people to make better choices for themselves

In order to achieve these goals, the BIT redesigns public services and uses ideas and findings from the behavioral science literature. They empirically test and trial these ideas first in small scale pilot programs before they are scaled up to evaluate whether a particular settings work or not. Nowadays, the BIT is considered a global organization which runs projects in over 15 countries, having offices in London, Sydney and New York (Behavioural Insights Team 2017).

3.3.1 Reciprocity and Fairness

The Behavioral Insights Team is engaged in several government policy areas around the world, consisting of e.g. health, education, governance or energy. One of the impressive examples of its activity and practical application of framing at the same time is the following experiment dealing with organ donation. Even though there is a lot of people registered in the NHS Organ Donor Register, the number is still not high enough to assure that organs are available for all those who need them. Roughly, three people die per day as a consequence of shortage of organs available. The problem is that there is a relatively large gap between intention and action as 9 out of 10 people consider organ donation to be important whilst fewer than 1 in 3 people is actually registered as a potential donor.

As an individual’s explicit consent is required in the UK in order to be placed upon the NHS Organ Donor Register, the BIT in cooperation with other several organizations addressed the issue by including framed messages
3. Framing in Field Experiments

and pictures when people applied for a driving license or renewed their vehicle tax. Once the applicant completed his or her transaction, the application web offered people to join the Register, including additional framed information about organ donation. The messages took various form; there were first, a control variant containing simple prompt to join, second a social norm appeal accompanied with two different pictures, third loss and gain framing, i.e. statement that 3 people die every day and the number of people you could potentially save as an organ donor, fourth appeal on individual’s fairness and reciprocity and finally recommendations to take rather action before intention. All of these interventions were chosen based on the evidence suggesting their positive effect on human behavior.

Interestingly, the most successful scheme was the Reciprocity one as the message: “If you needed an organ transplant, would you have one? If so please help others.” generated 1% more people registered as organ donors. This number might not seem to be especially high but if the Reciprocity message was used over the whole year, 96,000 additional registrations would be made on average in comparison with the control condition. Other interventions also significantly increased the number of registrations with only one exception. In contrast with the experimenters expectation, the social norm intervention stating that “Every day thousands of people who see this page decide to register” accompanied by the picture of a group of smiling people actually discouraged significant portion of people from registration. They conclude with hypothesis that the negative effect of the picture in social norm message was caused probably by using inappropriate stock photo.

Overall, the findings are important as they show how relatively small changes in particular public services may significantly affect outcomes. Further, they are promising improvement of the efficiency of these services, granting benefits for both sides - public providers as well as consumers. As it turned out, framing might not always perform as generally expected and therefore small trials are the feasible solution how to assess their performance (Behavioural Insights Team 2010).

3.3.2 Social Norm and Emotional Appeal

As written in another BIT study, almost 20% of charitable income in the United Kingdom is generated by donations through bequests, it might a fruitful area for behavioral research. Therefore the authors randomly assigned people calling
into a large national legal firm in order to arrange their wills into two groups. Callers from the control group were asked whether they have thought about leaving some money to charity (“weak ask”) whereas subjects assigned into the treatment group were apart from the “weak ask” confronted with a social norm, stating that leaving gift to charity is common, and an emotive prompt to think about charitable causes they are passionate about (“strong ask”).

**Weak Ask:** “Now that you’ve looked after your family and friends, I’d like to talk you about charity. Would you like to leave a charitable gift in your will?”

**Strong Ask:** “Now that you’ve looked after your family and friends, I’d like to talk to you about charity. Many of our customers like to leave a gift to charity in their will. Are there any charitable causes that you’re passionate about?”

Unsurprisingly, the probability of leaving money to charity turned out to be correlated with the form of interventions. The “weak ask” increased legacy giving by 5.4% compared to people who were not ask about making a charitable bequest at all. Furthermore, the “strong ask” persuaded approximately 11.4% callers more to legacy giving, i.e. taking it to three times the rate in the baseline sample. The experimenters also revealed that the intervention was much more effective for childless people as the “weak ask” raised the probability of legacy giving by 16 percentage points and the “strong ask” increased the probability of bequest by 31 percentage points (Sanders & Smith 2014).

### 3.4 Potential Concerns

Nevertheless, field experiments represent effective and promising way how to apply behavioral economics insights into everyday life, we should be aware of potential drawbacks that are crucial though. Most importantly, field environment contains a great variety of factors and characteristics that interact and influence each other. Thus, it might be challenging to correctly isolate the treatment effect from other forces and correctly interpret it. Although randomization process should assure that the observed effect is the true one at the same time, one should also take into account specific settings of the trial like e.g. individual experience, form of framing or specific aspects of the particular area of interest. This is well illustrated by the experiment of questioning an effect of default option on experienced subjects.
In the setting the environmental economists visiting a large international conference on environmental economics were asked whether they would wish to offset CO2 emissions if they were flying to the venue. The three different scenarios were designed: First, in the treatment “opt-in” the option “I do not want to compensate for my CO2 emissions” was pre-selected. Second, the treatment “opt-out” pre-selected the amount of money for emission compensation so if the participants did not want to contribute he or she had to switch the marked choice. Third, a control group faced an active decision with no default option.

Despite the large number of lab studies emphasizing the importance of default option or status quo, the experiments did not find any significance default bias. As they examined experienced subjects (i.e. participants had good knowledge about a good at hand) they argue that the effect of default option attenuate with level of experience. Therefore, policy makers should devote their attention to rather inexperienced subjects when applying default option policies as they can invoke substantially larger effects (Lofgren et al. 2012).
Chapter 4

Field Experiment: Testing
Enforcement Strategies in the Field

In this chapter we describe in detail the settings and results of a field experiment run by Fellner et al. (2013) which we asked student subject pool to predict.

4.1 Institutional Background

Being the main source of revenues, the collection of taxes and fees constitutes currently an important topic for the public sector entities, who provide a large variety of key services. Efficient collection of revenues could be especially challenging in settings under which a payer is asked to self-report information qualifying her for the payment. In the specific case of Austria each household that owns a TV is obliged by law to register and pay a monthly fee\(^1\) for national public television broadcasting. However an enforcement problem arises as the households are required to self-register while the public broadcasting is freely accessible to anyone without paying the fee.

The Austrian license fee system is managed by “Fee Info Service” (Gebühr Info Service, GIS for brevity) that is a subsidiary of the Austrian public broadcasting company. The main task of GIS is to collect and enforce TV license fees. GIS mails regularly throughout the year letters to potential evaders with information about their liability to pay the license and how to proceed in order to register and start paying. In order to identify residents that potentially evade paying, GIS dispose of its own database that is compared to the resi-

\(^1\)In 2005, the size of fee ranged from 206 to 263 euros, depending on a particular federal state (Fellner et al. 2013).
4. Field Experiment: Testing Enforcement Strategies in the Field

dence data. Therefore, households that have not registered a TV are treated as potential evaders. Once this group of interest is reached via letter, the License Fee Act requires those recipients to respond and provide correct information on why they do not pay license fees. Therefore, the variable of interest is not only a number households which register as a reaction to the mailings but also a number of households that respond to the letter and clarify their situation (Fellner et al. 2013).

4.1.1 Motivation

The traditional economic approach concerning the improvement of tax and fees collections relies on the insight that a rational taxpayer has incentives to avoid the taxes as long as the expected gains from evasion exceeds the costs of detection (Allingham & Sandmo 1972). As a result, the recommended way to address the issue of non-compliance is to strengthen the enforcement process and increase the probability of detection (improved monitoring, higher fines, prosecution, etc.). Nevertheless, such measures are usually costly to implement and in many cases it is questionable whether they really pay off. Behavioral economics disposes of the methods for the induction of behavioral changes that could provide similar effects as the traditional measures but with markedly lower expenses. Several studies on the tax-collection have shown that the manner in which certain information is presented in a regular mail can generate surprising effects\(^2\). Therefore, Fellner et al. (2013) address the enforcement problem in Austria in this manner.

4.2 Experiment Design

A standard GIS mailing to potential license fee evaders consists of a cover letter, an information sheet and a response and registration form with a postage prepaid envelope. The cover letter notifies the mailing recipient that since she has not registered any TV, she is suspicious of tax evasion and obliged to clarify her situation by responding to the letter within two weeks. The information sheet contains selected parts of the License Fee Act, i.e. information about the payment duty, size of the license fee and description of legal consequences in case of detection. For the purposes of the experiment, Fellner et al. (2013) manipulate the text of the cover letters (hereinafter referred to as “the letter” for

\(^2\)For more details see Chapter 3.
brevity), whilst everything else (the response form and the info sheet) remains the same. Naturally, the authors create also control group (T0) that do not receive any letter in order to construct a relevant counterpart to the mailing schemes.

### 4.2.1 Text Manipulations

First, as a baseline (T1) Fellner et al. (2013) use the standard letter that GIS sent in their previous mailing campaigns. Second, the authors present five text variants of the letter that differ in one or two paragraphs containing various information that addressed several distinct motives (or their combinations) of non-compliance to payment of the TV license fees. The five information manipulations are aimed at the most probable reasons for the avoidance of the fee payment.

The first variant (labeled “Threat”, T2) warns the recipients of what might happen if they are convicted of violation of the law, by pointing out GIS’ standard enforcement practice.\(^3\)

**Threat (T2)**

“*If you do not respond to this letter, a staff member of GIS will contact you in order to request information from you personally. If you refuse to provide information or if there is a well-founded suspicion that you provide disinformation, GIS is obligated to order an inquiry by the responsible federal authorities. Please keep in mind that in this case you may face legal consequences and considerable costs.*”

The second manipulation (“Social information”, T3) appeals on the social norm to comply to the laws and provides information about the actual level of tax fee compliance.\(^4\)

**Social information (T3)**

“*Do you actually know that almost all citizens comply with this legal duty? In*

---

\(^3\)For another example of “threat” manipulation see Subsection 3.2.3.  
\(^4\)For another example of “social norm” manipulation see Subsection 3.2.4.
fact, 94% - a vast majority of all households - have registered their broadcasting receivers.”

The third variant of the letter (“Moral appeal”, T5) incorporates moral appeal and points out that compliance behavior is ethically correct\(^5\).

**Moral appeal (T5)**

“Those who do not conscientiously register their broadcasting receivers not only violate the law, but also harm all honest households. Hence, registering is also a matter of fairness.”

The other two treatments are then combinations of the listed independent manipulations with the threat one (“Threat \times info”, T4 & “Threat \times moral”, T6).

### 4.3 Results

The following Figure 4.1 summarizes the results for the mailing treatments T1-T6, including the number of registrations, updates, and the overall response relative to the number of delivered letters. Obviously, there are substantial differences in the effectiveness of the individual treatments.

Concerning response rate the threat manipulation (T2) has generated the highest number of responses. The authors find a significant positive effect of the threat (T1 versus T2, \(p = 0.024\), T3 versus T4, \(p = 0.014\) and T5 versus T6, \(p = 0.000\)). Further, they indicate negative effects from the social information (T1 versus T3, \(p = 0.005\) and T2 versus T4, \(p = 0.008\)) and the moral appeal (T1 versus T5, \(p = 0.000\) and T2 versus T6, \(p = 0.010\)). Therefore, maybe surprisingly, the “threat” texting (T2) represents the only manipulation that is more effective than the baseline variant of the letter.

Similarly, the authors find a significant positive effect on the registration rate of the threat manipulation between treatments T1 and T2 (\(p = 0.034\)), T3 and T4 (\(p = 0.003\)) and between T5 and T6 (\(p = 0.020\)). On the other hand, no significant effects are detected either between social information and baseline (T3 versus T1, \(p = 0.420\)) and between moral appeal and baseline (T5 versus T1, \(p = 0.369\)). The same findings are valid when the two treatments are interacted with the threat (T4 versus T2, \(p = 0.948\) and T6 versus T2, \(p = 0.948\)).

\(^5\)For another example of “moral appeal” manipulation see Subsection 3.2.2.
Importantly, only 0.31% of individuals registered for license fees in the control group (T0) that did not receive any letter.

**Figure 4.1:** Mailing response within 50 days. Percentages are relative to the number of delivered mailings. Vertical lines indicate 95% confidence intervals.

*Source: Fellner et al. (2013).*
Chapter 5

Lab Experiment

We conducted a computerized laboratory experiment\(^1\) to test to which extent it is possible to forecast field experiment results\(^2\) in the lab. The results of this experiment will constitute new evidence on external validity of laboratory experiments and extend small but growing literature of predicting research results which we summarize in Section 2.2.

5.1 Experiment Design

Participants of the experiment (mainly university students) had to complete a task of making their own predictions of expected effects of the letter manipulations that were used in the field experiment conducted by Fellner et al. (2013). We assumed that the recruited subjects were not familiar with the field experiment in advance. The lab experiment consisted of six experimental sessions in four days that did not differ in treatments and subjects' decisions did not interact with each other and were not restricted by any time limits.

5.1.1 Task: Making Predictions

The participants performed a simple task designed in a similar way as in DellaVigna & Pope (2016a). First, the subjects got familiar with the system of financing public service television, i.e. how the potential evaders are addressed via mailings and that consequently they are obliged to respond to the letters and

---

\(^1\)To assure easy replicability of the experiment we used the Z-TREE program, Fischbacher (2007).

\(^2\)For the design of the predicted experiment see Chapter 4.
clarify their situation with license fees. Second, it was stated that there were two variables of interest:

- a total number of answers, i.e. whether recipient’s household responded to the letter (by updating their contact details or by filing a claim of possession of no TV)

- a total number of registrations, i.e. whether recipient’s household registered for regular payment of the TV license fees

In order to avoid possible confusion, we clearly emphasized that the total amount of answers clarifying the situations of the suspected recipients (response rate) included also the registrations (registration rate) and therefore it had to be always higher or equal to the total number of registrations.

Once the subjects were familiar with the context, they had some time to read carefully all the letters used in the field experiment that were sent by GIS. To make the work with letters easier we printed them out, organized them into a compact file and marked them with a label, stating a particular variant of the mailing\(^3\) (i.e. T1, T2, T3, T4, T5 and T6). Moreover, the manipulated paragraphs were highlighted just for the purpose of the lab experiment.

Afterward, participants were supposed to read a description that appeared on the monitor screen and introduced the task: “In this part of the experiment your task will be to predict how many % of households will respond to the mailing and how many % of households will register into a license fee database as a result of reaction to the received letter.” After reading the instructions, the subjects proceeded to a trial stage in which they could try how to control the program in order to make predictions and test how much money they would receive depending on the accuracy of their predictions\(^4\). The subjects were also informed that in the end of the experiment one of their predictions would be chosen at random and the payment would be calculated based on this particular prediction. This setting is incentive compatible as the participants pay attention to all the predictions equally (Grether & Plott 1979). Once everything was clear, the subjects proceeded to the main part of the experiment that is shown on the Figure 5.1.

In the main stage, the participants were asked to create forecasts about effects of the sent letters using slider scales. In order to provide the subjects with

---

\(^3\)See Appendix A for the letters.  
\(^4\)For more details on how we calculated payoff see Subsection 5.1.2
a relevant benchmark, the results of the baseline treatment (T1) were displayed at the top of the screen, using a fixed slider scale. To mitigate possible confusion and inconsistency of predictions, all the slider scales (both for response and registrations rates) were presented at once on one screen (i). Moreover, although the benchmark results were displayed in absolute numbers, the subjects made predictions relatively to the baseline treatment (ii). Importantly, we limited the scale in the sliders to a relative effect between -10% ad 10% (iii).

To summarize, in the first stage of the lab experiment the forecasters read a description of why and how GIS in Austria addresses problem of collection TV license fees, read the detailed instructions about their own task, saw the results for the baseline mailing (T1) and then made forecasts for the other 5 mailings, predicting both the response and registration rates.

5.1.2 Pay-off function

In order to elicit agent’s beliefs about a probabilistic event, we calculated the participants’ rewards based on the quadratic scoring rule (QSR), one of a family of so-called proper scoring rules (Winkler & Murphy 1968). Since the agent reports a prediction $\text{pred}$ about future event, the specific score employed is equal to $\text{max} – |\text{pred} – \text{real}|^2\alpha$, where $\text{max}$ is the maximum amount that the agent could earn, $\text{real}$ is the actual effect of the letter and $\alpha$ is a sensitivity parameter. Therefore, the more accurate a prediction is the more money is
5. Lab Experiment

paid to the participant. Nevertheless, the QSR maximizes an agent’s expected utility only if she is risk-neutral. Under risk aversion the marginal utility of the monetary payment to the subject confounds the effect of her beliefs, complicating belief elicitation. In order to control for risk-neutrality, we included a risk-related question (recommended and validated by Falk et al. (2016)) in the end of the experimental session.

5.1.3 Sample description

The experiment took place at the Laboratory of Experimental Economics (LEE) at the University of Economics in Prague and was attended by 94 participants in total. A majority of participants were students from Prague with the largest share of undergraduates (59.6%), followed by the graduates with a bachelor’s degree (25.5%) and a master’s degree (12.8%). The median age was 23 years and the most common field of study was economics and finance with a 42.5% share of the total sample. Participants were recruited using the online recruitment system ORSEE (Greiner 2015). The average payment was equal to 350 CZK, including a guaranteed show-up fee of 150 CZK and each experimental session lasted approximately 1 hour and 40 minutes.

5.1.4 Power Calculations

In order to determine the size of the experimental sample, we use the method of power calculations. A sample that is smaller than statistically required increases the probability of mistakenly concluding that there is no effect of the evaluated intervention when the opposite is actually true. We use G*Power, a power analysis program for statistical tests commonly used in social and behavioral research (Faul et al. 2007). Nevertheless, as noted in Duflo et al. (2007), power calculations in practice involve substantial guess work as one must have a prior idea of the mean and the variance of the outcome. Since our data are not normally distributed, we use a nonparametric Wilcoxon-Mann-Whitney test. We wish to achieve standard statistical power 80% with a significance level $\alpha = 0.05$ and a minimum effect size to which the test is sufficiently sensitive $d = 0.42$ (expecting mean group 1 (2) in letter forecasts equal to 44 (43) and $SD \sigma$ group 1 (2) equal to 2.50 (2.25)). According to Cohen (1977) this effect size is classified as a small to medium. Altogether with these parameters, we need to have 94 subjects in each group to achieve 80% statistical power.
5.2 Results

5.2.1 Response Rate

How does the actual response rate\(^5\) in the five experimental treatments compare to the forecasts made by the subject pool in the lab? Table 5.1 lists the individual treatments and reports the actual response rate in the treatment (Column 1) and the average forecast for that treatment made by the experimental pool (Column 2). Column 3 depicts the average absolute error, whilst Column 4 summarizes the mean of absolute individual errors. In Column 5 we indicate how many individuals outperform the mean prediction.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response Rate (1)</th>
<th>Mean Forecast (2)</th>
<th>Absolute Error, Mean Forecast (3)</th>
<th>Error, Indiv. Forecast (Mean and s.d.) (4)</th>
<th>Percent Individuals Outperforming Mean (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (T1)</td>
<td>43.090</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Threat (T2)</td>
<td>45.010</td>
<td>45.689</td>
<td>0.678</td>
<td>1.680*</td>
<td>20.680</td>
</tr>
<tr>
<td>Social Info (T3)</td>
<td>40.700</td>
<td>44.091</td>
<td>3.391</td>
<td>3.445*</td>
<td>53.200</td>
</tr>
<tr>
<td>Social Info + Threat (T4)</td>
<td>42.770</td>
<td>45.892</td>
<td>3.122</td>
<td>3.186*</td>
<td>50.000</td>
</tr>
<tr>
<td>Moral Appeal (T5)</td>
<td>38.820</td>
<td>44.108</td>
<td>5.288</td>
<td>5.288*</td>
<td>58.510</td>
</tr>
<tr>
<td>Moral Appeal + Threat (T6)</td>
<td>42.810</td>
<td>45.769</td>
<td>2.958</td>
<td>3.171*</td>
<td>51.060</td>
</tr>
<tr>
<td>Average Across the 5 Treatments (excl. Baseline)</td>
<td>42.022</td>
<td>45.110</td>
<td>3.087</td>
<td>3.354*</td>
<td>51.064</td>
</tr>
</tbody>
</table>

Notes: *the error is different from 0 at 0.01 significance level.

\(^5\)Throughout this chapter, both response and registration rate are indicated in %.
Table 5.2 provides information about statistical significance of differences between the treatments. Combined with Table 5.1 we can conclude that on average the forecasters expect all the treatments (T2-6) to be more effective than the baseline (T1). Further, they correctly anticipate that adding threat texting increases the response rate (T1,3,5 vs T2,4,6). Interestingly, for social info and moral manipulation they expect positive effect only when added into the baseline version of the letter (T1 vs T3,5) whilst no effect is expected when added into the threat one (T2 vs T4,6).

Table 5.2: Testing Differences between Treatments

<table>
<thead>
<tr>
<th>Baseline vs Forecasts</th>
<th>Threat</th>
<th>Social Info</th>
<th>Moral</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 vs T2*</td>
<td>T1 vs T2*</td>
<td>T1 vs T3*</td>
<td>T1 vs T5*</td>
</tr>
<tr>
<td>T1 vs T3*</td>
<td>T3 vs T4**</td>
<td>T2 vs T4</td>
<td>T2 vs T6</td>
</tr>
<tr>
<td>T1 vs T4*</td>
<td>T5 vs T6**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1 vs T5*</td>
<td>T1 vs T6*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: *the difference is significant at 0.01 significance level using Wilcoxon signed rank test for one sample (the largest p-value is 2.193 \times 10^{-8}, avoiding multiple comparison problem). **the difference is significant at 0.01 significance level using Wilcoxon rank sum test for two samples (the largest p-value is 9.867 \times 10^{-9}, avoiding multiple comparison problem).

Although Fellner et al. (2013) detect a significant negative effect of social info (T2 vs T4) and moral appeal (T2 vs T6) we find no statistically significant difference between these forecasts. The problem is, that we cannot be sure if either the forecasters indeed do not expect any difference or we lack a statistical power to detect the differences. Using G*Power program (Faul et al. 2007), we recalculate the optimal sample size with a new effect size that we can now directly obtain from the collected data. The effect size \(d\) is now equal to 0.097 for T2 vs T4 (0.037 for T2 vs T6) so we would need 1741 (12231) instead of 94 people to be 80\% sure that the null hypothesis is correctly rejected at 0.05 significance level. Figure 5.2 illustrates a non-linear relationship between the effect size and total sample size.
Figure 5.2: Power Calculations: Effect size $d$ and total sample size

Notes: *t* test - Means: Wilcoxon-Mann-Whitney test (two groups), Tails=Two, Parent distribution=Normal, Allocation ration $N_2/N_1$=1, $\alpha = 0.05$, $(1 - \beta) = 0.8$, produced with G*Power.

Figure 5.3 provides in graphical format the evidence on the accuracy of the average forecast. Each of the 6 points in the scatter plot represents a treatment, displaying the real response rate in the treatment (Column 1 in Table 5.1) and the average forecast (Column 2 in Table 5.1) on the $x$ and $y$ axis, respectively. Forecasts close to the 45 degree line indicate cases in which the average forecast is very close to the actual average performance (the Baseline treatment (T1) served as a benchmark for the forecasters).

Figure 5.3 shows that our subject pool predicts the response rate quite well; the correlation between the forecasts and the actual response rate is equal to 0.85 (Column 2 in Table 5.4). The black line captures the best interpolating line which has a slope of 0.242 (s.e. 0.244). The prediction is remarkably accurate for the Threat (T2) manipulation, with absolute error 0.678, but somewhat less accurate for the other three manipulations (T3, T4, T6), with absolute error $\approx 3.128$. The effect of Moral Appeal (T5) was predicted the most inaccurately as the absolute error of the mean forecast equals to 5.288. Across all 5 treatments, the average absolute error (Column 3 in Table 5.1) averages 3.087 percentage points, which is 7% of the average response rate across the treatments.
Figure 5.3: Wisdom-of-Crowds Accuracy: Average Response Rate and Average Forecast by Treatment

Notes: The continuous line indicates the OLS line fit across the 5 points, with estimate forecast = 34.589 (10.290) + 0.242 (0.244)*actual.

Another evidence on forecast accuracy can be found using a box plot. Figure 5.4 depicts median forecasts in the 5 treatments (thick lines placed in the boxes) and the actual response rates denoted by solid triangles. Apparently, except for the Threat (T2), the actual response rates do not fit between upper and lower quartiles of the predicted data, suggesting certain prediction inaccuracy in terms of absolute value. Nevertheless, Figure 5.4 captures relative trends between the treatments quite well. We now graphically visualize the previous finding that the experimental pool correctly predicts that adding “Threat” manipulation into the letters (T2, T4, T6) increases the response rates compared to the original letters (T1, T3, T5, respectively). Furthermore, the forecasters expect the treatments incorporating threat manipulations (T2, T4, T6) to be the only treatments that are distinctly more effective than the Baseline treatment (T1). However, the same trend is true only for the actual response rate of the treatment Threat (T2), as Social Info (T3) and Moral (T5) perform worse in comparison with the Baseline treatment (T1) in reality.
Figure 5.4: Box Plot: Average Response Rate and Average Forecast by Treatment

Notes: The colorized △ denotes the actual response rate. The Baseline treatment (T1) is added into the box plot in order to compare it to the other treatments.

Same trends captured by Figure 5.4 can be seen in Table 5.3 that summarizes the actual relative and predicted ordering of the treatments. Again, although the absolute predicted ranking of the treatments somewhat differs from the real one, the forecasters correctly anticipate increase of response rate in the treatments incorporating the “threat” manipulation (T2, T4, T6) compared to their counterparts (T1, T3, T5, respectively). Obviously, most confusion is associated with the first treatment\(^6\), with the difference in the actual and predicted rank equal to 3.426. The large majority (70%) of the forecasters consider the Baseline (T1) the less effective among treatments, whilst the opposite is true.

\(^6\) The Baseline (T1) served as a benchmark for forecasters as they predict the effects of the other 5 treatments relatively to it.
Table 5.3: Actual vs Predicted Rank Order: Response Rate

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Rank</th>
<th>Mean Forecasted Rank</th>
<th>Mean Forecasted Rank (abs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (T1)</td>
<td>2</td>
<td>5.426</td>
<td>6</td>
</tr>
<tr>
<td>Threat (T2)</td>
<td>1</td>
<td>2.628</td>
<td>3</td>
</tr>
<tr>
<td>Social info (T3)</td>
<td>5</td>
<td>4.138</td>
<td>5</td>
</tr>
<tr>
<td>Threat + Social info (T4)</td>
<td>4</td>
<td>2.021</td>
<td>1</td>
</tr>
<tr>
<td>Moral appeal (T5)</td>
<td>6</td>
<td>4.021</td>
<td>4</td>
</tr>
<tr>
<td>Threat + Moral appeal (T6)</td>
<td>3</td>
<td>2.064</td>
<td>2</td>
</tr>
</tbody>
</table>

So far, we have discussed forecasts made by a large number of subjects. Nevertheless, firms or policy-makers typically ask an individual expert for an opinion on future events since reaching a large number of forecasters might be expensive and complex task. Table 5.4 and Figure 5.5 provide information on the accuracy of individual forecasts using several measures. For the first measure (absolute error), we compute the absolute error in forecast by treatment and average across the 5 treatments. Similarly, we construct other accuracy measures, i.e. the squared error and the correlation and rank-order correlation between the 5 forecasts and the treatments.
Table 5.4: Accuracy of Forecasts versus Random Guesses: Response Rate

<table>
<thead>
<tr>
<th>Accuracy Measure</th>
<th>Average Accuracy (and s.d.) of Individual Forecasts</th>
<th>Accuracy of Mean Forecast (Wisdom of Crowds)</th>
<th>% Forecasters Doing Better Than Mean Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Absolute Error</td>
<td>3.354 (1.650)</td>
<td>3.087</td>
<td>51.064</td>
</tr>
<tr>
<td>Benchmark for Comparison Random Guess in 30-50%</td>
<td>1.978</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Squared Error</td>
<td>15.344 (10.849)</td>
<td>11.685</td>
<td>50.000</td>
</tr>
<tr>
<td>Benchmark for Comparison Random Guess in 30-50%</td>
<td>5.103</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td>0.511 (0.475)</td>
<td>0.849</td>
<td>22.340</td>
</tr>
<tr>
<td>Benchmark for Comparison Random Guess in 30-50%</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rank-Order Correlation</td>
<td>0.085 (0.333)</td>
<td>0.120</td>
<td>43.617</td>
</tr>
<tr>
<td>Benchmark for Comparison Random Guess in 30-50%</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.5a plots the cumulative distribution function of the absolute error for the 94 experimental subjects. The figure also displays the wisdom-of-crowds error (vertical red line) and benchmark for accuracy of prediction, i.e. random forecast between 30 and 55 percentage points (dotted blue line). The figure shows, that the accuracy of individuals is comparable with the accuracy of the average forecast. In other words 51% of individuals have higher accuracy than the average individual, and the average individual absolute error is 12% larger than the error of the average forecast (3.354 vs 3.087 in Column 1 and 2 in Table 5.4). The same trend is true for Figure 5.5b when using the negative mean squared error as the accuracy measure. Figures 5.5c and 5.5d, and Table 5.4 show the findings incorporating the two alternative measures of accuracy. When using (rank-order) correlation measures, majority of individuals do not do as well as the average individual, i.e. (57%) 78% of individuals.

For the random benchmarks, we draw 5,000 random forecasts from a uniform distribution in the specified range and then compute the average error over the 5,000 draws.
have lower accuracy than the average individual. This finding is referred as *wisdom of crowds*: the average over a crowd outperforms most individuals in the crowd. Moreover, as opposed to the absolute error measures (Figure 5.5a, 5.5b), although the large majority of individuals do not do as well as the average individual, they still outperform the random choice.

**Figure 5.5: Distribution of Accuracy Measures for Individuals: Response Rate**

Notes: The dotted blue line shows the counterfactual average absolute error assuming random forecasts between 30 and 55% response rate. The red line shows the absolute error for the average, as opposed to the individual, forecast.

In Figure 5.6 we decompose the negative mean absolute error accuracy measure by treatment. From the figure it is obvious that in the majority of treatments (T3, T4, T5, T6), wisdom-of-crowds effect is not present since the average over a crowd performs as well as half of the individuals in the crowd. Moreover, random guess is more accurate than the average forecast in these treatments. The only exception represents the Threat (T2) treatment in which the average forecast outperforms 80% of the individual forecasts as well as the random guess. Columns 4 and 5 in Table 5.1 summarizes the forecaster accuracy by treatment. Across treatments, 51 percent of subjects do better than the average as noted in Figure 5.6a.
Figure 5.6: Cumulative Distribution of Individuals’ Accuracy in Individual Treatments: Response Rate

Notes: The dotted blue line shows the counterfactual average absolute error assuming random forecasts between 30 and 55% response rate. The red line shows the absolute error for the average, as opposed to the individual, forecast.

So far we have treated the 94 subjects as interchangeable, and we have focused on the relationship between average and individual forecasts. Clearly, the ability to forecast future research might be affected by various factors. Although this research question does not constitute the main goal of this thesis, we now investigate whether the level of achieved education might have any affect on prediction accuracy. In the specific case we use the absolute (squared) error in forecast as the accuracy measure, we formulate a regression model as follows

\[ a_{i,t} = \alpha + \beta X_i + \eta_t + \epsilon_{i,t} \]

where an observation is a forecaster-treatment combination and \( a_{i,t} \) is a measure of accuracy for forecaster \( i \) and treatment \( t \). The independent variables are the expertise variables, i.e. we control for whether a student has obtained a university degree or not. The term \( \eta_t \) captures the treatment fixed effects whilst \( \epsilon_{i,t} \) measures the standard errors. The standard errors are clustered at the forecaster level to allow for correlation in errors across multiple forecasts.
by an individual. In the specific case we use the (rank-order) correlation between the forecast and response rate as the accuracy measure, we define a regression model as follows

\[ a_i = \alpha + \beta X_i + \epsilon_i \]

where the accuracy measure \( a_i \) is defined at the level of forecaster \( i \), as in contrast with the treatment-forecaster level.

Table 5.5 summarizes the results from the stated regression specifications according to various accuracy measures (Columns 1-4). The estimated coefficients are comparable to the accuracy measures reported in Table 5.4. Nonetheless, there is no statistically significant difference in accuracy across the groups (i.e. between undergraduates and graduates) according to these measures. This finding implies, that predictions made by graduate forecasters in this particular experiment are equally accurate as these made by undergraduate students.

Table 5.5: Graduates versus Undergraduates: Response Rate

<table>
<thead>
<tr>
<th>Dependent Variable (Measure of Accuracy)</th>
<th>Absolute Forecast Error Treat. t by Forec. i (1)</th>
<th>Squared Forecast Error Treat. t by Forec. i (2)</th>
<th>Rank-Order Correlation for Forecaster i (3)</th>
<th>Simple Correlation for Forecaster i (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (Undergraduates)</td>
<td>3.151* (0.184)</td>
<td>13.660* (1.558)</td>
<td>0.494* (0.064)</td>
<td>0.087 (0.045)</td>
</tr>
<tr>
<td>Graduates</td>
<td>0.049 (0.157)</td>
<td>0.196 (1.333)</td>
<td>0.040 (0.100)</td>
<td>-0.006 (0.070)</td>
</tr>
<tr>
<td>( N )</td>
<td>470</td>
<td>470</td>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td>( R )-squared</td>
<td>0.323</td>
<td>0.264</td>
<td>0.002</td>
<td>0.001</td>
</tr>
</tbody>
</table>

*significant at 0.01 level

Interestingly, DellaVigna & Pope (2016a) find the substantial differences in experimental findings' prediction accuracy between non-experts and experts. They define assistant, associate and full professors, post-docs, research scientists and PhD students as the experts whilst the non-experts consist of undergraduates, MBA students and online sample. Our experimental sample is probably too homogeneous and the fact that a forecaster has Bc, Mgr or PhD degree do not play any significant role in his or her prediction accuracy.
5. Lab Experiment

5.2.2 Registration Rate

Within this section, we report and discuss results concerning prediction of registration rate, i.e. how large share of addressed households register into the license fee database after they receive the letter. Table 5.6 again lists the individual treatments and reports the actual registration rate in each treatment (Column 1) and the average forecast for that treatment made by the forecasters (Column 2). Column 3 depicts the average absolute error, whilst Column 4 summarizes the mean of absolute individual errors. Indeed, the absolute differences in the registration rates across the treatments are less pronounced compared to the response rates (Table 5.1) but not in relative terms.

Table 5.6: Findings by Treatment: Registration Rate in Field Experiment and Forecasts (%)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Registration Rate</th>
<th>Mean Forecast</th>
<th>Absolute Error, Mean Forecast</th>
<th>Error, Indiv. Forecast (Mean and s.d.)</th>
<th>Percent Individuals Outperforming Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (T1)</td>
<td>8.620</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Threat (T2)</td>
<td>9.670</td>
<td>10.139</td>
<td>0.469</td>
<td>1.085* (1.180)</td>
<td>35.107</td>
</tr>
<tr>
<td>Social Info (T3)</td>
<td>8.230</td>
<td>8.991</td>
<td>0.761</td>
<td>1.042* (0.771)</td>
<td>42.553</td>
</tr>
<tr>
<td>Social Info + Threat (T4)</td>
<td>9.700</td>
<td>10.279</td>
<td>0.579</td>
<td>1.210* (1.291)</td>
<td>36.170</td>
</tr>
<tr>
<td>Moral Appeal (T5)</td>
<td>8.190</td>
<td>9.176</td>
<td>0.986</td>
<td>1.299* (1.096)</td>
<td>45.745</td>
</tr>
<tr>
<td>Moral Appeal + Threat (T6)</td>
<td>9.320</td>
<td>10.085</td>
<td>0.765</td>
<td>1.256* (1.282)</td>
<td>43.617</td>
</tr>
<tr>
<td>Average Across the 5 Treatments (excl. Baseline)</td>
<td>9.022</td>
<td>9.734</td>
<td>0.712</td>
<td>1.178* (0.903)</td>
<td>42.553</td>
</tr>
</tbody>
</table>

Notes: *the error is different from 0 at 0.01 significance level.

Table 5.6 combined with Table 5.7 suggest that on average the forecasters expect all the 5 treatments (T2-6) to be more effective than the baseline (T1).
They expect a positive effect of adding threat (T1,3,5 vs T2,4,6). The social info and moral appeal manipulations are considered to increase the registration rate only when added to the baseline (T1). Therefore, the relative trends for the registration rate forecasts are similar to the response rate, suggesting certain consistency in predictions.

Table 5.7: Forecasts: Testing Differences between Treatments

<table>
<thead>
<tr>
<th>Baseline vs Forecasts</th>
<th>Threat</th>
<th>Social Info</th>
<th>Moral</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 vs T2*</td>
<td>T1 vs T2*</td>
<td>T1 vs T3*</td>
<td>T1 vs T5*</td>
</tr>
<tr>
<td>T1 vs T3*</td>
<td>T3 vs T4**</td>
<td>T2 vs T4</td>
<td>T2 vs T6</td>
</tr>
<tr>
<td>T1 vs T4*</td>
<td>T5 vs T6**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1 vs T5*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1 vs T6*</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: *the difference is significant at 0.01 significance level using Wilcoxon signed rank test for one sample (the largest p-value is 5.013 \times 10^{-7}, avoiding multiple comparison problem). **the difference is significant at 0.01 significance level using Wilcoxon rank sum test for two samples (the largest p-value is 2.157 \times 10^{-6}, avoiding multiple comparison problem).

Nevertheless, Fellner et al. (2013) find no significant effect of social info and moral appeal (see Section 4.3) for registration rate. Thus, our forecasters are wrong saying there is an effect (T3,5) compared to the baseline (T1). If we want to claim they are correct for the threat (T2 vs T4 and T2 vs T6) we must again compute new optimal sample size, as we cannot be sure whether the null hypotheses about no difference between the treatments is not rejected wrongly due to the lack of statistical power in our experimental design. Using G*Power we would need 2189 (13979) forecasters in each group to detect effect size \( d = 0.087 \) (0.034) for T2 vs T4 (T2 vs T6).

Figure 5.7 provides in graphical format the evidence on the accuracy of the average forecast. Each of the 6 points in the scatter plot represents a treatment, displaying the real registration rate in the treatment (Column 1 in Table 5.6) and the average forecast (Column 2 in Table 5.6) on the x and y axis, respectively. Forecasts close to the 45 degree line indicate cases in which the average forecast is very close to the actual average performance (the Baseline treatment (T1) served as a benchmark for the forecasters). Again, the forecasters overestimate the absolute effect of the rated treatments. Interestingly, the forecasted registration rates are almost equally distant from the 45 degree line, suggesting high average accuracy in prediction of relative differences between the 5 treatments. This is confirmed when we use the correlation as the accuracy
measure of mean forecast. The correlation is equal to stunning 0.983 (Column 2 in Table 5.9), meaning that the forecasters change their relative predictions appropriately with the varying treatments.

**Figure 5.7:** Wisdom-of-Crowds Accuracy: Average Registration Rate and Average Forecast by Treatment

Notes: The continuous line indicates the OLS line fit across the 5 points, with estimate forecast = 1.566 (2.093) + 0.890 (0.233)*actual.

Another evidence on forecast accuracy can be found in the following box plot. Figure 5.8 depicts median forecasts in the 5 treatments (thick lines placed in the boxes) and the actual registration rates denoted by solid triangles. In contrast with the box plot for response rate (Figure 5.4), the actual reponse rates fit between upper and lower quartiles of the predicted data for the 3 treatments (T2, T4, T6). Consistently, the forecasters again anticipate an increase in registration rates when “threat” manipulation is added into the original letter (T2, T4, T6) compared to the respective letters without the “Threat” (T1, T3, T5).
Figure 5.8: Box Plot: Average Registration Rate and Average Forecast by Treatment

Table 5.8 summarizes the actual relative and predicted ordering of the treatments. Again, although the absolute predicted ranking of the treatments somewhat differs from the real one, the forecasters correctly anticipate increase of response rate in the treatments incorporating the “threat” manipulation (T2, T4, T6) compared to their counterparts (T1, T3, T5, respectively). Compared to Table 5.3, the forecasters are consistent in the rank ordering of the treatments. Since the real rank of the treatments differs for response and registration rate, the forecasters now correctly predict the absolute rank of two treatments with social info manipulation, i.e. T3 (5) and T4 (4). Again, they mark the baseline (T1) as the weakest treatment but now with a minor error than in the response rate ranking (1.351 versus 3.246).

Table 5.9 provides information on the accuracy of individual forecasts using several measures and compare them to the randomly generated forecasts. Again, we use the mean absolute error, the squared error and the correlation and rank-order correlation between the 5 forecasts and the treatments that are listed in the rows of Table 5.9. If we compare the average of individual fore-
casts (Column 1) with the mean forecast (Column 2) it is obvious, that the mean forecast is more accurate for all the accuracy measures. Furthermore, the present wisdom-of-crowds effect assures, that the average over a crowd outperforms most individuals in the crowd. The accuracy is especially stunning for the correlation measure as only 2 persons out of 94 people are more accurate than the mean forecast. Although the large majority of individuals do not do as well as the average forecaster, they still outperform random choice. This finding is at odds with its response rate counterpart, where the random choice performs better than the absolute error of the average for negative error measures. The results summarized in Table 5.4 are graphically displayed in Figure 5.9.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Rank</th>
<th>Mean Forecasted Rank</th>
<th>Mean Forecasted Rank (abs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (T1)</td>
<td>4</td>
<td>5.351</td>
<td>6</td>
</tr>
<tr>
<td>Threat (T2)</td>
<td>2</td>
<td>2.553</td>
<td>3</td>
</tr>
<tr>
<td>Social info (T3)</td>
<td>5</td>
<td>4.245</td>
<td>5</td>
</tr>
<tr>
<td>Threat + Social info (T4)</td>
<td>1</td>
<td>1.968</td>
<td>1</td>
</tr>
<tr>
<td>Moral appeal (T5)</td>
<td>6</td>
<td>3.947</td>
<td>4</td>
</tr>
<tr>
<td>Threat + Moral appeal (T6)</td>
<td>3</td>
<td>2.149</td>
<td>2</td>
</tr>
</tbody>
</table>
### Table 5.9: Accuracy of Forecasts versus Random Guesses: Registration Rate

<table>
<thead>
<tr>
<th>Accuracy Measure</th>
<th>Average Accuracy (and s.d.) of Individual Forecasts (1)</th>
<th>Accuracy of Mean Forecast (Wisdom of Crowds) (2)</th>
<th>% Forecasters Doing Better Than Mean Forecast (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean Absolute Error</strong>&lt;br&gt; <em>Benchmark for Comparison</em></td>
<td>1.178 (0.903)</td>
<td>0.712</td>
<td>42.553</td>
</tr>
<tr>
<td>Random Guess in 0-20%</td>
<td>1.364</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mean Squared Error</strong>&lt;br&gt; <em>Benchmark for Comparison</em></td>
<td>2.684 (4.173)</td>
<td>0.538</td>
<td>36.170</td>
</tr>
<tr>
<td>Random Guess in 0-20%</td>
<td>2.785</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong>&lt;br&gt; <em>Benchmark for Comparison</em></td>
<td>0.614 (0.424)</td>
<td>0.983</td>
<td>2.128</td>
</tr>
<tr>
<td>Random Guess in 0-20%</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rank-Order Correlation</strong>&lt;br&gt; <em>Benchmark for Comparison</em></td>
<td>0.520 (0.345)</td>
<td>0.718</td>
<td>36.170</td>
</tr>
<tr>
<td>Random Guess in 0-20%</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Figure 5.10 we again decompose the negative mean absolute error accuracy measure by treatment. From the figure it is obvious that in the majority of treatments (T3, T4, T5, T6), a mild wisdom-of-crowds effect is present since the average over a crowd performs better than approximately 60% of the individuals in the crowd. The stronger wisdom-of-crowds effect (80%) is present for T2 treatment. Moreover, random guess is more accurate than the average forecast in almost all the treatments (T2, T3, T5, T6). Across treatments, 42 percent of individual subjects do better than the average as noted in Figure 5.10a.
Figure 5.9: Distribution of Accuracy Measures for Individuals: Registration Rate

Notes: The dotted blue line shows the counterfactual average absolute error assuming random forecasts between 0 and 20% registration rate. The red line shows the absolute error for the average, as opposed to the individual, forecast.

Similar to Table 5.5, Table 5.10 summarizes the results from the stated regression specifications\textsuperscript{8} according to various accuracy measures (Columns 1-4). The estimated coefficients are comparable to the accuracy measures reported in Table 5.9. Nonetheless, there is no statistically significant difference in accuracy across the groups (i.e. between undergraduates and graduates) according to these measures. This finding implies, that predictions made by graduate forecasters in this particular experiment are equally accurate as these made by undergraduate students.

\textsuperscript{8}See Section 5.2.1 for more details.
5. Lab Experiment

Figure 5.10: Cumulative Distribution of Individuals’ Accuracy in Individual Treatments: Registration Rate

Notes: The dotted blue line shows the counterfactual average absolute error assuming random forecasts between 0 and 20% registration rate. The red line shows the absolute error for the average, as opposed to the individual, forecast.

Table 5.10: Graduates versus Undergraduates: Registration Rate

<table>
<thead>
<tr>
<th>Dependent Variable (Measure of Accuracy)</th>
<th>Absolute Forecast Error Treat. t by Forec. i (1)</th>
<th>Squared Forecast Error Treat. t by Forec. i (2)</th>
<th>Rank-Order Correlation for Forecaster i (3)</th>
<th>Simple Correlation for Forecaster i (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (Undergraduates)</td>
<td>1.207* (0.125)</td>
<td>3.003* (0.629)</td>
<td>0.575 (0.056)</td>
<td>0.485 (0.046)</td>
</tr>
<tr>
<td>Graduates</td>
<td>0.123 (0.107)</td>
<td>0.496 (0.538)</td>
<td>0.096 (0.089)</td>
<td>0.087 (0.072)</td>
</tr>
<tr>
<td>N</td>
<td>470</td>
<td>470</td>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.010</td>
<td>0.012</td>
<td>0.013</td>
<td>0.016</td>
</tr>
</tbody>
</table>

*significant at 0.01 level
5.2.3 Summary

If we measure prediction accuracy of *response rate* in the absolute numbers (the mean absolute error, the mean squared error), the forecasters’ accuracy vary with the treatments. Moreover, the accuracy of individuals is comparable to the accuracy of the average forecast, i.e. no wisdom-of-crowds effect is present. The only exception represents the Threat (T2) treatment with the mean smallest absolute error equal to 0.678 and 80% wisdom-of-crowds effect (Column 3 and 5 in Table 5.1). We obtain better results with usage of the correlation between the forecast and the response rate across the treatments (0.849, Column 2 in Table 5.4), meaning that the forecasters correctly anticipate the direction in which the response rates vary among the 5 treatments.

Interestingly, when we measure prediction accuracy of *registration rates* in the absolute numbers, it might seem that the forecasters perform very well as their absolute error of the mean forecast does not exceed 1 percent point (Column 3 in Table 5.6). Nevertheless, if this error is averaged across the 5 treatments it represents almost 8% of the absolute error of the mean forecast and therefore it has the same magnitude as the one associated with *response rate*. Furthermore, we detect that majority of individuals do not do as well as the average individual, suggesting an omnipresent wisdom-of-crowds effect. This finding is especially striking when we use correlation accuracy measure, as the accuracy of mean forecast reaches 0.983, outperforming 98 percent of individual forecasts (Column 2 and 3 in Table 5.9).

Most importantly, the forecasters correctly predict that the treatments containing threat (T2, T4, T6) do better than their original counterparts (T1, T3, T5). On average, the forecasters anticipate the baseline treatment (T1) to be the less effective compared to the other treatments for both response and registration rate. More specifically, they expect that adding social info (T3) or moral appeal (T5) into the baseline letter should persuade more people to reply or register. In reality however, this is not true as the treatments containing solely moral appeal (T5) and social info (T3) perform worse (response rate) or the same (registration rate) as the baseline. We assume this can have two reasons, i.e. either the *floor effect* of the baseline treatment was too pronounced or the subjects really consider social and moral manipulations to be that effective. Although based on the evidence we present in Chapter 3 one could expect the latter, it might be interesting to replicate the study and clearly emphasize, that the individual treatments might perform *worse* in comparison with the
baseline one. Nevertheless, this discrepancy reminds us that one should always carefully consider context of the experiment and specific characteristics of the particular setting when making forecasts. Furthermore, there is no statistically significant difference between the forecasts of “threat” and “social info” (T2 vs T4) and “threat” and “moral appeal” (T2 vs T6) for both response and registration rate which is again at odds with Fellner et al. (2013) who find a negative effect for response rate. In order to be sure that we cannot reject the null hypotheses about no difference between (T2 vs T4,6) just because of lacking statistical power we would need to increase our total experimental sample size to 1741 (12231) subjects. To summarize, the forecasters correctly predict the positive effect of the threat but they are not able to anticipate effects of social info and moral on average.

Finally, we do not find any statistically significant differences in the prediction accuracy based on the level of achieved education, i.e. differences between graduates and undergraduates. Nevertheless, this analysis is not the ultimate goal of this thesis as it would probably require to incorporate more heterogeneous subject pool.
Chapter 6

Conclusion

In this thesis we have focused on the problem of forecasting the results of a field experiment in a laboratory setting. Laboratory generated forecasts might serve as first indicators of effectiveness of the individual treatments incorporated in field studies and might play an important role in final experimental design. Our goal has been to scrutinize to which extent such forecasts can predict the field results, investigating the external validity of laboratory experiments at the same time. Therefore, we design a novel lab experiment in which the recruited subjects (forecasters) are first introduced to the context and design of the specific field experiment and consequently they make predictions about its results. To establish prediction accuracy we compare the collected data with the actual results, challenging the often criticized external validity of laboratory experiments.

Field experiments organized in cooperation with public and private entities are currently experiencing a rapid development as they present an advantageous way to go for both groups, i.e. economists and organizations. Nonetheless, to prepare and carry out a large field experiment might turn out to be time consuming, not to mention that inappropriately designed treatments can cause a significant loss not only in financial terms but also in the reputation of an organization and trustworthiness of experimental economists. Although a typical behavioral manipulation should not be expensive to employ from the definition, when applied to tens of thousands of subjects it might come at cost. Therefore, it might be beneficial to forecast results of field studies in a simple lab environment in order to obtain first insight about magnitude of variables of interest and get better idea about effectiveness of the intended treatments.

Firstly, we have designed and conducted a lab experiment with 94 experi-
mental subjects, mainly university students. Their task was to predict results of a field experiment carried out by Fellner et al. (2013) in Austria in 2005. In the study the authors vary the texting of letters sent to households that were suspected of evading TV license fees, focusing on response and registration rates invoked by the mailings. Similarly, our laboratory subject pool made forecasts on both response and registration rate, after being introduced to the institutional background of TV license fee collection. We used one of the treatments (Baseline (T1)) as a benchmark, providing the forecasters with the actual result of this treatment to get better idea about the real rates and refine their predictions.

Secondly, we have analyzed the collected data with several accuracy measures, arriving to the following results. First, the average forecast accuracy among the 94 forecasters vary with the treatments. The most accurate forecast is the “threat” treatment (T2) with absolute error of 0.678 (0.469) for response (registration) rate. Across the 5 treatments the average absolute deviation between the average forecast and the outcome is about 7 (8) percent of the average score. Second, we document differences in accuracy between the average forecast and individual forecasts. Depending on a type of accuracy measure and individual treatments we conclude that the average forecast is always more accurate or at least comparable to the mean individual forecast, proving the presence of wisdom-of-crowds effect. Third, we reveal that forecasters correctly predict that adding threat texting into the letters (T2, T4, T6) improves the effectiveness of their threat-free counterparts (T1, T3, T5), i.e. the positive effect of the threat. Nonetheless, the forecasters are wrong when they anticipate that social info and moral appeal (T3, T5) persuade more households to reply or register compared to the baseline treatment (T1). In order to reliably evaluate the relative effect of social info, moral appeal and threat (T2 vs T4, T6) forecasts we would need higher statistical power and therefore substantially increase the subject pool. Our last finding is that the level of achieved education does not have any effect on forecasting accuracy as undergraduates perform as well as graduates.

This thesis constitutes new evidence in small, yet growing literature focused on forecasts of research findings. Clearly, there is still a lot of work to be done but we consider the topic highly relevant, as it opens promising opportunities for future research. It turns out that that we can elicit relevant forecasts even from inexperienced subject when provided with appropriate information.
Bibliography


Bibliography


ORTMANN, A. (2003): “Field experiments in economics; some methodological
caveats.” Technical report, Mimeo, CERGE.

OTTAVIANI, M. & P. N. SORENSON (2007): “Aggregation of information and
beliefs in prediction markets.” Technical report, Finance Research Unit at
University of Copenhagen.


Concept, design and implementation for a sales forecasting problem.” Tech-
nical report, California Institute of Technology Social Science.

PLOTT, C. R. & S. SUNDER (1988): “Rational expectations and the aggrega-
tion of diverse information in laboratory security markets.” Econometrica:


Sense? How well do experts and lay-people do at predicting the findings of

Determinants of Charitable Bequests.” Technical report, The Centre for
Market and Public Organisation.

SCHRAM, A. (2005): “Artificiality: The tension between internal and external
validity in economic experiments.” Journal of Economic Methodology 12(2):
pp. 225–237.

SERVAN-SCHREIBER, E., J. WOLFERS, D. M. PENNOCK, & B. GALEBACH


Appendix A

Text of Cover Letters
Dear Mr. Michael Gruber,

You listen to radio, you watch TV? Then you are aware of the program variety offered by Austrian Public Broadcasting. The provision of these services, however, requires funding. Therefore, everybody who owns a radio or a TV has to pay license fees. It is the task of GIS Gebühren Info Service GmbH to ensure that all TV and radio consumers pay these fees.

Our database does not show a registration of TV or radio equipment at your address. This can have several reasons:

- We may have made a mistake in our database and you are already registered at GIS. In this case, we apologize in advance.
- Your registration data may have changed, e.g., due to a move or a name change (marriage), and our computer system cannot match the data with your registration.
- You may not hold a radio or a TV at this address and therefore do not have to register anything.
- Maybe you have just forgotten to register your TV or radio.

We are legally obliged to clarify this issue and kindly ask you to answer our questions — even if you have already registered at GIS. On the back of this letter you will find a response form. Please fill in this form and send it back within the next 14 days.

We thank you for your cooperation. If you require further information, please call our service hotline at 0810 00 10 80 (Monday to Friday, 8.00am to 9.00pm, Saturday from 9.00am to 5.00pm) or visit our web page at www.orf-gis.at.

Kind regards,

GIS-Team
GIS GMBH, 1051 Wien, Postfach 1000

Dear Mr. Michael Gruber
Linzstraße 1
1100 Wien

Dear Mr. Michael Gruber,

You listen to radio, you watch TV? Then you are aware of the program variety offered by Austrian Public Broadcasting. The provision of these services, however, requires funding. Therefore, everybody who owns a radio or a TV has to pay license fees. It is the task of GIS Gebühren Info Service GmbH to ensure that all TV and radio consumers pay these fees.

Our database does not show a registration of TV or radio equipment at your address. This can have several reasons:

- We may have made a mistake in our database and you are already registered at GIS. In this case, we apologize in advance.
- Your registration data may have changed, e.g., due to a move or a name change (marriage), and our computer system cannot match the data with your registration.
- You may not hold a radio or a TV at this address and therefore do not have to register anything.
- Maybe you have just forgotten to register your TV or radio.

We are legally obliged to clarify this issue and kindly ask you to answer our questions – even if you have already registered at GIS. On the back of this letter you will find a response form. Please fill in this form and send it back within the next 14 days.

If you do not respond to this letter, a staff member of GIS will contact you in order to request information from you personally. If you refuse to provide information or if there is a well-founded suspicion that you provide disinformation, GIS is obligated to order an inquiry by the responsible federal authorities. Please keep in mind that in this case you may face legal consequences and considerable costs.

We thank you for your cooperation. If you require further information, please call our service hotline at 0810 00 10 80 (Monday to Friday, 8.00am to 9.00pm, Saturday from 9.00am to 5.00pm) or visit our web page at www.orf-gis.at.

Kind regards,
GIS-Team
GIS GMBH, 1051 Wien, Postfach 1000

Dear Mr.
Michael Gruber
Linzstraße 1
1100 Wien

Dear Mr. Michael Gruber,

You listen to radio, you watch TV? Then you are aware of the program variety offered by Austrian Public Broadcasting. The provision of these services, however, requires funding. Therefore, everybody who owns a radio or a TV has to pay license fees. It is the task of GIS Gebühren Info Service GmbH to ensure that all TV and radio consumers pay these fees.

Do you actually know that almost all citizens comply with this legal duty? In fact, 94% – a vast majority of all households – have registered their broadcasting receivers.

Our database does not show a registration of TV or radio equipment at your address. This can have several reasons:

- We may have made a mistake in our database and you are already registered at GIS. In this case, we apologize in advance.
- Your registration data may have changed, e.g., due to a move or a name change (marriage), and our computer system cannot match the data with your registration.
- You may not hold a radio or a TV at this address and therefore do not have to register anything.
- Maybe you have just forgotten to register your TV or radio.

We are legally obliged to clarify this issue and kindly ask you to answer our questions – even if you have already registered at GIS. On the back of this letter you will find a response form. Please fill in this form and send it back within the next 14 days.

We thank you for your cooperation. If you require further information, please call our service hotline at 0810 00 10 80 (Monday to Friday, 8.00am to 9.00pm, Saturday from 9.00am to 5.00pm) or visit our web page at www.orf-gis.at.

Kind regards,

GIS-Team
Dear Mr. Michael Gruber,

You listen to radio, you watch TV? Then you are aware of the program variety offered by Austrian Public Broadcasting. The provision of these services, however, requires funding. Therefore, everybody who owns a radio or a TV has to pay license fees. It is the task of GIS Gebühren Info Service GmbH to ensure that all TV and radio consumers pay these fees.

Do you actually know that almost all citizens comply with this legal duty? In fact, 94% – a vast majority of all households – have registered their broadcasting receivers.

Our database does not show a registration of TV or radio equipment at your address. This can have several reasons:

- We may have made a mistake in our database and you are already registered at GIS. In this case, we apologize in advance.
- Your registration data may have changed, e.g., due to a move or a name change (marriage), and our computer system cannot match the data with your registration.
- You may not hold a radio or a TV at this address and therefore do not have to register anything.
- Maybe you have just forgotten to register your TV or radio.

We are legally obliged to clarify this issue and kindly ask you to answer our questions – even if you have already registered at GIS. On the back of this letter you will find a response form. Please fill in this form and send it back within the next 14 days.

If you do not respond to this letter, a staff member of GIS will contact you in order to request information from you personally. If you refuse to provide information or if there is a well-founded suspicion that you provide disinformation, GIS is obligated to order an inquiry by the responsible federal authorities. Please keep in mind that in this case you may face legal consequences and considerable costs.

We thank you for your cooperation. If you require further information, please call our service hotline at 0810 00 10 80 (Monday to Friday, 8.00am to 9.00pm, Saturday from 9.00am to 5.00pm) or visit our web page at www.orf-gis.at.

Kind regards,

GIS-Team
Dear Mr. Michael Gruber,

You listen to radio, you watch TV? Then you are aware of the program variety offered by Austrian Public Broadcasting. The provision of these services, however, requires funding. Therefore, everybody who owns a radio or a TV has to pay license fees. It is the task of GIS Gebühren Info Service GmbH to ensure that all TV and radio consumers pay these fees.

Those who do not conscientiously register their broadcasting receivers not only violate the law, but also harm all honest households. Hence, registering is also a matter of fairness.

Our database does not show a registration of TV or radio equipment at your address. This can have several reasons:

- We may have made a mistake in our database and you are already registered at GIS. In this case, we apologize in advance.
- Your registration data may have changed, e.g., due to a move or a name change (marriage), and our computer system cannot match the data with your registration.
- You may not hold a radio or a TV at this address and therefore do not have to register anything.
- Maybe you have just forgotten to register your TV or radio.

We are legally obliged to clarify this issue and kindly ask you to answer our questions – even if you have already registered at GIS. On the back of this letter you will find a response form. Please fill in this form and send it back within the next 14 days.

We thank you for your cooperation. If you require further information, please call our service hotline at 0810 00 10 80 (Monday to Friday, 8.00am to 9.00pm, Saturday from 9.00am to 5.00pm) or visit our web page at www.orf-gis.at.

Kind regards,

GIS-Team
Dear Mr. Michael Gruber,

You listen to radio, you watch TV? Then you are aware of the program variety offered by Austrian Public Broadcasting. The provision of these services, however, requires funding. Therefore, everybody who owns a radio or a TV has to pay license fees. It is the task of GIS Gebühren Info Service GmbH to ensure that all TV and radio consumers pay these fees.

Those who do not conscientiously register their broadcasting receivers not only violate the law, but also harm all honest households. Hence, registering is also a matter of fairness.

Our database does not show a registration of TV or radio equipment at your address. This can have several reasons:

- We may have made a mistake in our database and you are already registered at GIS. In this case, we apologize in advance.
- Your registration data may have changed, e.g., due to a move or a name change (marriage), and our computer system cannot match the data with your registration.
- You may not hold a radio or a TV at this address and therefore do not have to register anything.
- Maybe you have just forgotten to register your TV or radio.

We are legally obliged to clarify this issue and kindly ask you to answer our questions – even if you have already registered at GIS. On the back of this letter you will find a response form. Please fill in this form and send it back within the next 14 days.

If you do not respond to this letter, a staff member of GIS will contact you in order to request information from you personally. If you refuse to provide information or if there is a well-founded suspicion that you provide disinformation, GIS is obligated to order an inquiry by the responsible federal authorities. Please keep in mind that in this case you may face legal consequences and considerable costs.

We thank you for your cooperation. If you require further information, please call our service hotline at 0810 00 10 80 (Monday to Friday, 8.00am to 9.00pm, Saturday from 9.00am to 5.00pm) or visit our web page at www.orf-gis.at.

Kind regards,

GIS-Team