



UvA-DARE (Digital Academic Repository)

Behavioral economics and the public sector

Weber, M.G.

Publication date

2015

Document Version

Final published version

[Link to publication](#)

Citation for published version (APA):

Weber, M. G. (2015). *Behavioral economics and the public sector*.

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.



Behavioral Economics and the Public Sector

Matthias Weber

This thesis consists of four essays dealing with topics that are relevant for the public sector. The essays cover diverse issues of economics partly overlapping with political science. The topics reach from the taxation of labor over monetary policy to preferences over voting institutions. Throughout this thesis it is, in contrast to classical economics, not assumed that humans are necessarily fully rational. Once full rationality is no longer assumed, experiments become an important tool to learn about human behavior. Consequently, most of the work in this thesis makes use of economic experiments.

Matthias Weber studied mathematics at the University of Freiburg (Germany) and economics at the Tinbergen Institute (Amsterdam, Netherlands). From 2011 to 2015 he conducted his PhD-research at the Center for Experimental Economics and Political Decision Making (CREED) at the University of Amsterdam. His research interests lie in the fields of public economics, macroeconomics, and political economics. Within these fields he approaches most problems from a behavioral or experimental perspective. Currently, he works at the Bank of Lithuania and Vilnius University (Lithuania).

Behavioral Economics and the Public Sector
Matthias Weber



Universiteit van Amsterdam



BEHAVIORAL ECONOMICS AND
THE PUBLIC SECTOR

ISBN 978 90 5170 668 0

Cover photo: Peter Lober

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. 631 of the Tinbergen Institute Research Series, established through cooperation between Thela Thesis and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Behavioral Economics and the Public Sector

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D. C. van den Boom
ten overstaan van een door het
college voor promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 17 november 2015, te 10:00 uur
door

Matthias Gerhard Weber

geboren te Schwäbisch Hall, Duitsland

Promotiecommissie:

Promotor: Prof. A. J. H. C. Schram Universiteit van Amsterdam

Overige leden: Prof. J. Duffy University of California, Irvine
Prof. J.-F. Laslier Centre National de la Recherche
Scientifique, Paris
Prof. J. H. Sonnemans Universiteit van Amsterdam
Dr. J. van de Ven Universiteit van Amsterdam
Prof. F. A. A. M. van Winden Universiteit van Amsterdam

Acknowledgments

I would like to thank everyone who contributed to the development of this thesis.

First and foremost, thanks go to Arthur Schram who has been a great advisor. Arthur is a co-author of the first essay, but he also advised me on all other chapters. His support and guidance in the last years have proved extremely valuable.

I would also like to thank the co-authors of the second essay in this thesis, Cars Hommes and Domenico Massaro. The collaboration was very pleasant and productive.

Furthermore, I would like to thank the whole of CREED, which provided me with an exceptionally fruitful research environment during the time I worked on this thesis. CREEDers provided valuable feedback in different seminars, in various discussions over lunch or coffee, in the hallway, and even inside the offices.

Additional thanks for comments on and suggestions for one or more of the projects in this thesis at any stage go to Ghazala Azmat, Manfred Holler, Maximilian Hoyer, Aaron Kamm, Sascha Kurz, Boris van Leeuwen, Nicola Maaser, Stefan Napel, Hannu Nurmi, Pedro Robalo, Rei Sayag, David Smerdon, Joep Sonnemans, Johannes Vatter, and Jeroen van de Ven.

In the last years I have spent some time on research visits at other universities. I would like to thank Jordi Brandts for hosting me at the Universitat Autònoma de Barcelona and Uri Gneezy for hosting me at the University of California in San Diego.

While working on this thesis I also spent some time on other scientific projects. These include an article on the future of power index research with Sascha Kurz, Nicola Maaser, and Stefan Napel (Kurz et al., 2015), a paper on a statistical method to incorporate network information into regression settings with Martin Schumacher and Harald Binder (Weber et al., 2014, based on Weber, 2009), and an experiment on bond market behavior with John Duffy and Arthur, which is still work in progress. I would like to thank my co-authors for their efforts in these collaborations.

Last but not least I would like to thank my family, my friends, and everyone else who has supported me during the last years.

Contents

1	Introduction	1
2	The Non-Equivalence of Labor Market Taxes: A Real-Effort Experiment	7
2.1	Introduction	7
2.2	Related Literature	11
2.3	Experiment	13
2.3.1	Treatments	13
2.3.2	Course of Events	15
2.4	Hypotheses and Mechanisms	21
2.5	Results	24
2.5.1	Public Sector Size Preference	24
2.5.2	Subjective Well-Being	25
2.5.3	Labor Supply and Job Performance	27
2.6	Discussion	32
	Appendix 2.A Instructions, Test Questions, and Questionnaire	35
	2.A.1 Instructions and Test Questions	35
	2.A.2 Questionnaire	50
	Appendix 2.B Screenshots	52
	Appendix 2.C Sample Composition and Gender Effects	56
	2.C.1 Sample Composition	56
	2.C.2 An Explorative Analysis of Gender Effects	56
3	Monetary Policy under Behavioral Expectations: Theory and Experiment	59
3.1	Introduction	59
3.2	Theory	61
3.2.1	Macroeconomic Model	61
3.2.2	A Behavioral Model of Expectation Formation	62
3.2.3	Monetary Policy, Inflation, and Output Gap	64

3.3	Experiment	67
3.3.1	Treatments and Hypotheses	68
3.3.2	Course of Events and Implementation	69
3.3.3	Results	71
3.4	Concluding Remarks	74
Appendix 3.A	Additional Graphs from Simulations of the Macroeconomic Model	75
Appendix 3.B	Instructions in the Experiment	79
Appendix 3.C	Additional Graphs of the Experimental Data and Screenshot . . .	83
4	Two-Tier Voting: Measuring Inequality and Specifying the Inverse Power Problem	93
4.1	Introduction	93
4.2	One Theoretical Concept: Penrose's Square Root Rule	95
4.3	Measuring the Inequality of Voting Systems	96
4.4	The Inverse Power Problem	98
4.4.1	Error Terms Based on Voting Power on the Group Level	98
4.4.2	An Error Term Based on Normalized Indirect Voting Power	99
4.4.3	Using the Coefficient of Variation to Specify the Inverse Problem	99
4.5	Illustrations	101
4.5.1	First Example: Mean-Preserving Spread	101
4.5.2	Second Example: Shift of the Distribution	103
4.5.3	Third Example: Comparing the Inequality of Voting Systems across Different Populations I	104
4.5.4	Fourth Example: Comparing the Inequality of Voting Systems across Different Populations II	104
4.6	Concluding Remarks	105
5	Choosing the Rules: Preferences over Voting Systems in Assemblies of Representatives	107
5.1	Introduction	107
5.2	Equalizing Voting Power and Maximizing Utility	110
5.2.1	Penrose's Square Root Rule and Proportional Shapley-Shubik Power	111
5.2.2	Choosing Voting Systems According to Expected Utility Theory	112
5.3	Experimental Design and Procedures	114
5.3.1	Illustration of Voting Systems	114
5.3.2	Treatments and Overview	115
5.3.3	Voting Procedures and Payments	117

5.3.4	Relationship between Theory and Treatments	119
5.3.5	Decision Situations	120
5.4	Results	122
5.4.1	Subjects Prefer Rule II (Proportional Shapley-Shubik Power) over Rule I (Penrose's Square Root Rule)	123
5.4.2	Subjects React to Changes in the Payoff Structure	125
5.4.3	Subjects Choose Differently When They Are in Front of the Veil (to Their Own Group's Benefit)	127
5.5	Discussion	130
Appendix 5.A	Instructions, Test Questions, and Questionnaire	132
5.A.1	Screenshots of Instructions and Test Questions, Treatment <i>BI</i>	132
5.A.2	Instruction Differences in the In Front of the Veil Treatments	145
5.A.3	Instruction Differences in the Group Aligned Preference Treatments	147
Appendix 5.B	Decision Blocks and Additional Graphs and Data	150
5.B.1	Properties of the Decision Blocks	150
5.B.2	Further Information on the Selection of the Decision Situations Used in the Experiment	157
5.B.3	Additional Graphs and Data	158
	Bibliography	163
	Summary	177
	Samenvatting (Summary in Dutch)	181

List of Tables

2.1	2×2 design	14
2.2	Illustration of wages, taxes, and earnings in the experiment	15
2.3	Overview and tests, public sector size preference	25
2.4	Regression results for public sector size preference	26
2.5	Overview and tests, subjective well-being	26
2.6	Regression results for subjective well-being	27
2.7	Overview and tests, labor supply at the extensive margin	28
2.8	Regression results for labor supply at the extensive margin, measure 1	29
2.9	Regression results for labor supply at the extensive margin, measure 2	29
2.10	Overview and tests, labor supply at the intensive margin and job performance	30
2.11	Regression results for labor supply at the intensive margin	31
2.12	Regression results for job performance	31
2.13	Sample composition	56
2.14	Results split up according to gender	57
3.1	Set of heuristics	64
5.1	2×2 between subjects design	117
5.2	Voting systems in the decision blocks	121
5.3	Efficiency of the voting systems in treatment <i>BA</i>	122
5.4	Differences in choices behind and in front of the veil	129
5.5	Properties, decision block 1	151
5.6	Properties, decision block 2	152
5.7	Properties, decision block 3	153
5.8	Properties, decision block 4	154
5.9	Properties, decision block 5	155
5.10	Properties, decision block 6	156
5.11	Choice proportion data	160
5.12	Counts of the most preferred voting system per block	161

List of Figures

2.1	Timeline of the experiment	16
2.2	Screenshot during a work round (taken in treatment IN)	17
2.3	Subjective well-being self-assessment manikin	18
2.4	Slider for the elicitation of public sector size preference	21
2.5	Screenshot during a work round	53
2.6	Screenshot during the self-assessment with the SAM-V-9	54
2.7	Screenshot of the slider; public sector size preference	55
3.1	Inflation volatility as a function of ϕ_y for the rational and the behavioral model	65
3.2	Output gap volatility as a function of ϕ_y for the rational and the behavioral model	66
3.3	Fractions of heuristics used for behavioral expectations of output gap and inflation	68
3.4	Hypotheses about inflation volatility	69
3.5	Screenshot	70
3.6	Realized inflation for all groups in both treatments	72
3.7	Realized output gap for all groups in both treatments	72
3.8	Empirical distribution functions of inflation volatility	73
3.9	Inflation volatility in the rational model as a function of ϕ_y for different values of ϕ_π	75
3.10	Inflation volatility in the behavioral model as a function of ϕ_y for different values of ϕ_π	75
3.11	Inflation volatility in the behavioral model for different starting values	77
3.12	Inflation volatility in the behavioral model for different starting values	78
3.13	Relation score and forecast error (not labeled in the instructions)	82
3.14	Realizations and forecasts of inflation and output gap ($T1$, groups 1 – 6)	84
3.15	Realizations and forecasts of inflation and output gap ($T1$, groups 7 – 12)	85
3.16	Realizations and forecasts of inflation and output gap ($T1$, groups 13 – 17)	86

3.17	Realizations and forecasts of inflation and output gap ($T2$, groups 1 – 6) . . .	87
3.18	Realizations and forecasts of inflation and output gap ($T2$, groups 7 – 12) . . .	88
3.19	Realizations and forecasts of inflation and output gap ($T2$, groups 13 – 15) . . .	89
3.20	Realizations and forecasts of inflation and output gap (excluded groups) . . .	90
3.21	Screenshot	91
5.1	Screenshot in front of the veil	116
5.2	Screenshot behind the veil	118
5.3	Proportion of participants predominantly choosing one system	123
5.4	Participants' choices	124
5.5	Differences in choices and expected payoffs in treatment BA	126
5.6	Difference in differences BI to BA , choices, and expected payoffs	127
5.7	Choices in front of the veil for the smallest and largest groups	128
5.8	Welcome screen	133
5.9	Instructions voting 1	134
5.10	Instructions voting 2	135
5.11	Instructions voting 3	136
5.12	Instructions voting 4	137
5.13	Test questions voting 1	138
5.14	Test questions voting 2	139
5.15	Continuation screen	140
5.16	Instructions choices and payoffs 1, treatments BI and FI	141
5.17	Instructions choices and payoffs 2, treatments BI and BA	142
5.18	Test questions choices and payoffs	143
5.19	Summary screen, treatments BI and FI	144
5.20	Instructions choices and payoffs 2, treatments FI , and FA	146
5.21	Instructions choices and payoffs 1, treatments BA and FA	148
5.22	Summary screen, treatments BA and FA	149
5.23	Proportion of participants predominantly choosing one system	158
5.24	Overview of the data	159

Chapter 1

Introduction

As the title implies, this thesis is concerned with behavioral economics and with the public sector. Behavioral economics is the part of economics that aims at investigating humans' actual behavior and at analyzing the consequences of actual human behavior for economic outcomes. 'Actual behavior' stands in contrast to assuming a world of fully rational economic agents. For scholars from disciplines other than economics it may be surprising, but assuming that people act rationally without room for mistakes or variations in the perception of the world around them is what economists have been almost exclusively assuming over many decades. Evidence is abundant that actual human behavior is often very different from the classical economic paradigm, however (see Tversky and Kahneman, 1974, Grether and Plott, 1979, Kahneman et al., 1991, and Conlisk, 1996, among many others). This does not mean that human behavior can never be rational, there are certainly situations in which rationality describes human behavior well, however it should not be taken for granted independent of the environment, the tasks at hand, and the complexity of a situation. This thesis does not assume full rationality and is thus part of the field of behavioral economics.

As soon as it is accepted that humans need not necessarily be rational, economic experiments become an important tool to learn about human behavior. It may be surprising to people who are unfamiliar with behavioral and experimental economics that it is possible to gain insights about economics by gathering people in a computer laboratory and paying them to make decisions; even more so when the studies are in fields such as public finance or macroeconomics. However, this is indeed the case if the experiments are well-designed. One of the largest advantages of experiments is that they allow one to establish causality. This is extremely difficult using empirical field data, e.g. because of confounding factors or reverse causality.¹ Laboratory experimentation allows us to keep everything constant except

¹Consider a large tax reform as an example. Assuming that we have good data, we can observe how behavior is different after the tax reform. However, it is very difficult to say which part of the tax reform causes behavior to change. The behavior might also just change because other things are changing in the world that

for the treatment intervention. Thus, changes in behavior must be caused by the treatment intervention. In theory, different results between the treatments could be due to chance and the heterogeneity of participants, but randomization and statistical analysis make it highly unlikely that something else than the treatment intervention drives the results. Laboratory experiments furthermore allow us to design a setting which exactly implements the characteristics of the the question at hand. This means that one can make a setting that is exactly like the setting that one wants to analyze rather than relying on observational data that are from only similar settings (if such data are available at all).

Probably the most common concern about laboratory experiments is their external validity. Thus, the question is to what extent the experimental results carry over to the outside world. This is an important question (somewhat similar to the question concerning the extent to which economic models have a meaning for the real world). In most well-conducted experiments, it is possible to induce similar trade-offs or characteristics in the laboratory as in the outside world. A theory that only depends on trade-offs and/or characteristics that can be induced in the laboratory should hold in the laboratory just as much as in different situations with these trade-offs outside of the laboratory.² If such a theory does not find support in the laboratory, it seems unlikely that it will hold in the more general setting outside of the laboratory – economic theories are often applied to a wide variety of situations, in rich and poor countries alike, in all climates, for construction workers and for teachers, in work environments and during private interactions, etc. Thus, as far as the main ingredients of economic models are implemented in the laboratory, these findings can be important for our understanding of human behavior also outside of the laboratory.

For the reasons outlined above, economic experiments play an important role in this thesis. For more extensive discussions of the use of experiments in economics, see for example Schram (2005), Falk and Heckman (2009), or Charness and Kuhn (2011).

Not taking full rationality for granted (and similarly making use of experimental methods) is a common element of the essays in this thesis, which is independent of the topic studied. This thesis is composed of four essays which are largely independent of one another. These essays do have in common that all of the topics studied are relevant for the

have nothing to do with the tax reform. Or the tax reform was only implemented because of a change in the world that is also correlated to the change in behavior. Often, it turns out to be very difficult to disentangle the different effects.

²As one example, consider individuals' labor supply decisions. In the outside world people have very different work experiences, depending for example on the job specification, the work environment, and individual characteristics. Nevertheless, economists assume that labor supply decisions mainly depend on a trade-off between the disutility of working and the utility of the monetary compensation (the extent of disutility from working and the extent of utility gained from working can in general differ). This trade-off can easily be induced in the laboratory. Thus, a theory that only depends on this trade-off and for example on the perception of the monetary compensation should hold in the laboratory just as anywhere else.

public sector. ‘The public sector’ is of course something very broad, comprising many different institutions and having many functions. This breadth is reflected by the topics in this thesis. They reach from the taxation of labor over the conduct of monetary policy to the perception of voting institutions.

Chapter 2 is concerned with labor market taxation. In almost all countries labor is taxed in one way or another, often even in different ways within one country. A large share of public revenues consists of labor taxes and these taxes are paid by a large fraction of the workforce worldwide. Studying labor taxes is thus very important. In this chapter, we³ investigate whether and how the liability side of a tax on labor matters. It is a classic economic result that it does not matter under fully rationality whether these taxes are paid by employers or by employees. Not only does the liability side of the tax not matter for who bears the burden of the tax, but it also has no influence on individuals’ labor supply or voting decisions. This does not necessarily need to be the case if people are not fully rational, however. We design a laboratory experiment in which we can manipulate whether the tax is born by employers or employees. We do this in a simple setting where we only change the framing of the tax (thus, we really give liability side equivalence the best possible chance to hold – tasks and payments are very salient and the difference in the liability side is exclusively implemented through framing). Then we investigate whether political economic preferences, labor supply, and subjective well-being are different under these distinct taxes. There are of course reasons to assume that they are indeed different; we also describe the behavioral mechanisms that we expect to govern individuals’ reactions to the two taxes in this chapter.

Note that this experiment provides a good example of how experiments can be used to establish causality. There is only a small treatment intervention (changing the framing of a tax) which is the only thing that can be responsible for systematic differences in behavior between the treatments. This study also nicely shows how a relatively simple policy change can cause behavioral change in multiple dimensions. Thus, focusing on only one dimension when giving policy advice can turn out to be too narrow (when thinking about labor taxes, labor supply is clearly the most obvious dimension for economists).

Chapter 3 is concerned with aggregate macroeconomic behavior and its implications for monetary policy. As monetary policy is one of the main determinants of inflation and unemployment, its conduct has enormous impacts and studying it is clearly important. In this chapter, we replace the common assumption of rational expectations in a New Keynesian framework by the assumption that expectations are formed according to a heuristics switch-

³Chapter 2 is based on joint work with Arthur Schram (Weber and Schram, 2015) and Chapter 3 is based on joint work with Cars Hommes and Domenico Massaro (Hommes et al., 2015). Consequently, when writing about the work of these two chapters I use the first person plural while I use the first person singular in the remainder.

ing model. This behavioral model of expectation formation assumes that individuals make their expectations using relatively simple heuristics and switch between these heuristics depending on their past performance. That is, individuals use heuristics more often that have been closer to the actual outcomes in the past. We study how the economy behaves under the rational and the behavioral models with a special focus on price stability, more precisely on inflation volatility. According to the behavioral macroeconomic model, it can be beneficial for the central bank to target both inflation and the output gap even if the central bank is ultimately only interested in price stability. This is different in the analogous model based on rational expectations. These opposing theoretical predictions are then tested in a learning-to-forecast experiment in the laboratory. The only difference between the treatments lies in the parameters of the monetary policy equation simulating the behavior of the central bank.

This chapter again provides a nice example of establishing causality through experimentation. The only thing that is different between the treatments in the experiment is the law of motion of the time series, not even the instructions differ. This chapter furthermore illustrates nicely how macroeconomics can benefit from the use of laboratory experiments.

Chapters 4 and 5 are concerned with political institutions, more precisely with voting in assemblies of representatives. Such assembly voting takes place when there are different groups sending out one representative each to a voting assembly. This way of collective decision making is used by a wide variety of institutions, including the Council of the European Union, UN General Assembly, German Bundesrat, ECB, and thousands of boards of directors and professional and non-professional associations. Studying it is thus certainly important.

Chapter 4 is a bit shorter than the other chapters and it is not making use of laboratory experimentation. In it, I ask how one should measure the extent of inequality in such voting systems. I propose to use the coefficient of variation to do so and show how this relates to another problem. This other problem is how to find voting systems that approximate equal indirect voting power as well as possible (equal indirect voting power refers to the theoretical ideal of all citizens from all groups being equally likely to influence the outcome of the decision in the assembly). This chapter nicely shows the relation between specifying the inverse power problem with a particular inequality measure and specifying it with more classic objective functions.

Chapter 5 is about preferences over voting system for assemblies of representatives. There is an abundant normative theoretical literature examining the distribution of voting power that voting systems ‘should’ follow. However, nobody has ever investigated which voting systems people actually prefer. This is important for the legitimacy and acceptance of voting institutions. It can furthermore have an influence on peoples’ behavior as individuals react to the institutions and procedures in place. In this chapter, I show the findings of a labo-

ratory experiment in which participants choose voting systems behind the veil of ignorance, that is when they do not know which group they will be in. As a control, also participants' choices in front of the veil are observed, that is when they do know which group they will be in. The experimental design is such that it is possible to observe the extent to which participants choose voting systems designed according to the most prominent social choice rules.

Note the novelty of the approach taken in this chapter. Hundreds of normative papers have been written on assembly voting. Yet, the question of which voting systems people actually prefer has, to the best of my knowledge, never been studied. Furthermore, this chapter provides a nice example of how laboratory experiments allow us to investigate topics that would be extremely hard to investigate in the outside world. There, it would for example be impossible to examine choices behind the veil of ignorance as people always know which group(s) they belong to.

Chapter 2

The Non-Equivalence of Labor Market Taxes: A Real-Effort Experiment*

2.1 Introduction

Traditional public finance assumes full rationality when analyzing the economic impact of taxes. Under this assumption classic results on tax incidence can straightforwardly be derived, such as liability-side equivalence (LSE). In the words of Joseph Stiglitz:

“It makes no difference whether a tax is imposed on the suppliers of a factor or commodity rather than on the consumers. (...) Taxes induce changes in relative prices, and it is this market response that determines who bears the tax.”
(Stiglitz, 2000, p. 514)

Full rationality is a questionable assumption if it aims to describe human behavior in the real world, however. Since at least Simon (1955) the evidence of bounded rationality in economic decisions has accumulated (Conlisk, 1996). For the study of tax incidence, bounded rationality introduces the relevance of issues like tax perception, framing, myopia, or time inconsistency (Bassi, 2010). Assuming full rationality may therefore have far-reaching consequences. Consider, for example, Stiglitz’ assessment. The underlying assumption in the assertion that relative prices determine tax incidence via market responses is that individuals correctly perceive taxes and respond to them in a utility-maximizing manner. If bounded rationality affects either perception or response, prices no longer fulfill this role and LSE is no longer obvious.⁴

*This chapter is based on joint work with Arthur Schram (Weber and Schram, 2015).

⁴Another assumption commonly made in the traditional public finance literature is that individuals have self-regarding preferences. Numerous papers in behavioral and experimental economics have shown the prominence of other-regarding preferences, however (for a survey, see Cooper and Kagel 2009). If other-regardness

In this chapter, we study the perception and the behavioral responses to distinct labor taxes that are equivalent in the traditional sense. This is important since labor taxes play a major role in all modern economies and many tax policies are still based on the lessons obtained in traditional public finance. The recently emerged field dubbed ‘behavioral public finance’ (e.g. McCaffery and Slemrod, 2006, Mullainathan et al., 2012) intends – among other things – to mend this lack of an empirically sound basis for economic policies. This is the field to which this research hopes to contribute.⁵ We test one of the prerequisites for tax equivalence to hold. Labor tax equivalence follows from a rational perception of distinct taxes in combination with utility maximizing choices and market forces. We specifically study the first element, i.e., whether the framing of otherwise equivalent taxes affects behavior. Rational perception (which in our case is the absence of a framing effect) is a necessary, but not sufficient condition for tax equivalence to hold. If framing effects are observed, this provides direct evidence against tax equivalence. Moreover, while most economists think of tax equivalence only in terms of market prices and quantities, there are more ways in which taxes can be equivalent under full rationality. With boundedly rational agents, this equivalence could for example be violated if the distinct taxes induce individuals to prefer the provision of different quantities of a tax-financed public good or if the taxes lead to different levels of subjective well-being.

Two ways of taxing labor prevail around the world. One is an income tax levied on employees, the other a payroll tax levied on employers.⁶ Note that these two types of taxes exist side by side in many countries. This is somewhat surprising from a full rationality point of view (at least, in the absence of market frictions), because under equivalence one would expect the tax to be chosen that minimizes collection and compliance costs. The reason for the co-existence is possibly that people perceive the two taxes differently and react to them in different ways. This is what this chapter investigates.

There are many ways in which such taxes may differentially affect people. Here, we list three. First, there may be strong effects on individual political preferences. If perceptions

takes processes into account (as opposed to being outcome-based), distinct taxes may not be valued equally. Moreover, if bounded rationality affects the perception of or the response to taxes, other-regarding preferences may inflate the differences.

⁵A small part of the traditional public finance literature allows for failure of LSE in the labor market due to market frictions. We are not interested here in studying labor market frictions, but in perceptions of and reactions to taxes that are not necessarily rational. Therefore, we study these taxes in a setting where they are by design equivalent under full-rationality, i.e. in a setting without frictions. We will not mention this equivalence every time we compare the taxes.

⁶Employer payroll taxes often take the form of contributions (for example to social security or health care). Legally, there is a difference between taxes and contributions as in the latter case, employees usually receive an entitlement that they do not receive with a tax. We treat the terms as equivalent here. We use the term ‘income tax’ for a tax (or contribution) on the employees’ (supply) side of the labor market and the term ‘employer payroll tax’ for a tax (or contribution) on the employers’ (demand) side.

vary across taxes of how much of the public sector is financed by distinct groups in society, opinions on the preferred size of the public sector (or welfare state) are likely to vary as well. This could directly influence voting decisions. Intuitively, it might explain why right-wing politicians tend to favor duties levied on the employees' side while left-wing politicians tend to favor duties levied on the employers' side. Second, many economists are interested in the effects of policies on some index of well-being (representing individual utility or its aggregate, social welfare). Subjective well-being is an obvious first measure of the consequences of policies, including taxes.⁷ People simply might be happier under some tax regimes than under others. In the end, a third effect of labor market taxes is probably to many economists the most obvious. This is their effect on labor supply decisions and on job performance (or total output produced). Individuals may decide to work more or less under one tax regime than under another, either at the intensive margin (hours worked) or at the extensive margin (labor market participation). It may occur, for example, that high gross wages induce people to accept jobs that they would not accept after careful consideration of post-tax income.

We consider the effects of differential perception of theoretically equivalent labor market taxes on each of these three dimensions: political-economic preferences, subjective well-being, and labor supply (together with performance). Simultaneously considering multiple dimensions is important because even simple policy changes – such as the change in the liability side of a labor tax (even with enforced equilibrium wages) – can affect welfare in many ways. It is far from trivial to derive policy implications from these complex effects, but ignoring important dimensions when judging the welfare properties of a policy change can lead to incorrect policy recommendations and possibly very costly mistakes. Our three dimensions represent the three main categories through which individuals may be affected: preferences, well-being, and economic decisions.

For our purpose, observational field data are usually ill-suited, because it is generally difficult to disentangle the numerous effects stemming from broad tax reforms. It is also often impossible to filter out the causes of observed effects (e.g., differences could stem from market frictions or from differences in perception) and moreover, counterfactuals are missing in such data. In addition, field experiments on taxes are almost impossible to implement as governments are highly unlikely to agree to implement a treatment design including proper controls, because not all citizens would be treated equally.⁸

This leaves laboratory experiments as a natural choice to investigate the questions at hand. Even if other empirical methods were feasible, for various reasons such experiments

⁷For an overview of the literature on the measurement of subjective well-being, see Kahneman and Krueger (2006). For discussions on using such measures for welfare comparisons, see Anand and Van Hees (2006), Schokkaert (2007), and Ferrer-i Carbonell (2013).

⁸Nevertheless, there are a few examples of field experiments on taxation that have been successfully implemented (e.g., the New-Jersey/Pennsylvania Negative Income Tax experiments; see Robins, 1985).

would still be a preferred way to investigate this issue. For one thing, the laboratory allows one to provide a setting that is most favorable for liability side equivalence to hold. All tasks, payoffs, and taxes are more salient and more directly related to decisions than is typically the case outside of the laboratory. Furthermore, institutional frictions are absent and laboratory control allows one to make the taxes equivalent by design instead of being equivalent only in general equilibrium. As a consequence, a lack of LSE in the laboratory – where it is given its best shot – would raise serious doubts about its validity outside of the laboratory. In addition, in a careful experimental design one can systematically vary the environment in which the taxes are implemented, which allows one to test the sensitivity of LSE to such changes. For example, we will distinguish between an environment in which proceeds are lost, and one where tax revenues are used to produce a public good. Finally, the laboratory provides the opportunity to directly measure the effects of taxation. In particular, it allows us to directly measure subjects' preferences for the size of the public sector, their subjective well-being and their labor supply responses.

In sum, we examine in a laboratory experiment with human subjects and monetary incentives whether people react differently to an incentive scheme depicting an income tax than to one reflecting a payroll tax levied on employers. By design, both duties are absolutely equivalent under full rationality. To increase the external validity of our laboratory environment, the experiment will require real effort by subjects to earn an income (that may subsequently be taxed). Our distinction between an environment where proceeds are lost and one when a public good is produced allows us to isolate the effects on LSE of (perceived) returns from taxation.

This research is primarily empirical in the sense that it carefully establishes in a controlled environment whether framing effects exist that contradict LSE. In addition, however, we distinguish between three mechanisms – gross wage illusion, tax loss effect, and warm glow – that would explain such framing effects and discuss how these mechanisms interact to predict the experimental results we observe.

Our results suggest that differences in the way the two taxes are perceived affect behavior in each of the dimensions that we distinguish between. More specifically, in our experiment employer-side taxes lead to (1) workers preferring a larger public sector; (2) higher subjective well-being of the workers; and (3) lower labor supply (at least at the extensive margin); all in comparison to the case where the taxes are levied on workers' gross income. Each of these effects indicates that it matters who actually transfers the taxes to the government.

This chapter intends to add to the literature in the following ways. It is the first to investigate how levying a labor tax on either the employers' or the employees' side of the labor market influences individuals' preferences concerning the size of the public sector. To elicit these preferences in the laboratory, we introduce a novel, incentive compatible approach.

This research is also the first to investigate the effects of the liability side of a labor tax on subjective well-being. In addition, it provides further evidence on the effects of the liability side of a labor tax on labor supply at the extensive and intensive margin and on job performance, being the first to do so in an environment that mimics an employer-employee relationship.

The organization of this chapter is as follows. Section 2.2 reviews the related literature and Section 2.3 presents the experimental design and procedures. This is followed by Section 2.4, which contains the hypotheses to be tested and the psychological mechanisms we expect to be at work. Section 2.5 presents the results. In Section 2.6 we discuss the results and their implications.

2.2 Related Literature

To our knowledge there are no previous studies investigating the effects of the liability side of a labor tax on political-economic preferences. Nor do we know of any study of the effects of such taxes on subjective well-being. There are, however, other studies examining labor supply or job performance under such taxes. A part of this literature is theoretical.⁹ In this short overview we focus on the related empirical – especially experimental – literature.

There is not much non-experimental empirical research that is closely related. A notable exception is Lehmann et al. (2013) who investigate how gross earnings change when income tax rates or payroll tax rates change using recent French data. They find that gross labor earnings respond to changes in the marginal income tax rate while they do not respond to changes in the payroll tax rate, thus rejecting LSE. The authors suggest that this might be due to differential effects of these tax changes on labor supply. Using data from the Netherlands Muysken et al. (1999) present evidence that a larger part of taxes is shifted if they are levied on the employees' side rather than on the employers' side. Holm et al. (1994) use Finnish data in an empirical application of a monopoly union wage determination model and find that increasing the payroll tax rate has a negative effect while increasing the income tax rate has a positive effect on wages. Using Greek data from the early nineteen-nineties, Saez et al. (2012) find that upper income earners do not respond to increases in payroll taxes concerning their labor supply decisions, neither on the intensive nor on the extensive margin.

⁹There are different approaches to modeling the labor market and thus the impact of labor market taxes. Most prominent are the competitive labor market approach (see Atkinson and Stiglitz, 1980), the efficiency wage theory (Shapiro and Stiglitz, 1984), search and matching models (see Pissarides, 2000) and union bargaining models (see Oswald, 1985). Most of this literature does not allow for liability side non-equivalence. Such non-equivalence can arise in exceptional cases via market frictions, however (e.g. Koskela and Schöb, 1999, Picard and Toulemonde, 2001, Rasmussen, 1997). In none of these approaches there is any non-equivalence due to tax perception.

There is a (limited) experimental literature on the effects of taxes on labor supply. We know of no such study implementing an employer-employee relationship in the laboratory. Hayashi et al. (2013) experimentally study distinct income tax schemes (specifically, no tax, a flat tax, a progressive tax and a wage subsidy) that are equivalent under full rationality. Their results show that labor supply differs across treatments. Their most robust finding is that subjects choose to supply less labor in the wage subsidy treatment than in the others. Fochmann et al. (2010b) investigate whether the gross wage has an influence on the labor supply decisions of the participants. They find that participants choose to work longer and harder when their gross wage is higher (holding net wages constant). They refer to their finding as ‘gross wage illusion’.¹⁰ Djanali and Sheehan-Connor (2012) find that experimental subjects work harder for the same net wage when the gross wage is higher and the tax proceeds go to a non-profit organization, in line with various social preference theories. Finally, in a field experiment that does not involve taxes, Hossain and List (2012) show that workers in a Chinese factory respond differentially to distinct framing of productivity bonuses. All of these results hint at possible effects that tax framing may have in the labor market.

Other papers examine liability side equivalence of taxes in situations resembling a more general buyer-seller environment. Sausgruber and Tyran (2005) study the perception and effects of direct and indirect taxes. They find that the tax burden associated with an indirect tax is underestimated, which is not the case with a corresponding direct tax. Their study also shows that this can lead to voting for inefficiently high redistribution. Experience seems to weaken this effect, however. Sausgruber and Tyran (2011) add to their previous research by showing that while experience is an effective de-biasing mechanism, pre-vote deliberation about tax regimes is not. Riedl and Tyran (2005) investigate gift exchange markets. Their results support LSE. Finally, Cox et al. (2012) examine tax incidence in double auction and posted offer markets and find that LSE does not hold. All in all, the experimental work on LSE in various environments provides mixed results.

Considering tax perception more generally, many studies report seemingly irrational behavior by laboratory participants. An excellent survey is presented in Fochmann et al. (2010a). An example of this literature is De Bartolome (1995), who shows that many people mistakenly use the average tax rate instead of the marginal tax rate when making investment decisions. Fochmann et al. (2012) study how investment decisions change with the framing of taxes. Their experimental results show that the possibility to deduct losses from an income tax leads to significantly riskier investments (their treatments are equivalent under full rationality). Blumkin et al. (2012) find that experimental subjects in the laboratory supply less labor under an income tax than under a consumption tax (when both taxes are equiva-

¹⁰Fochmann and Weimann (2013) elaborate on Fochmann et al. (2010b) and explain these findings by tax salience.

lent under full rationality). Ullmann and Watrin (2008) conduct experiments showing that people are more likely to evade taxes in a consumption tax environment than in an income tax environment. Such ‘irrationality’ carries over to the field. Chetty et al. (2009) report on a field experiment suggesting that consumers react differently when sales taxes are already included in the price tag than when they are not included in it. These authors also deserve credit for making the concept of tax salience prominent. Finkelstein (2009) provides evidence that tolls become less salient when collected electronically and that drivers’ behavior then becomes less elastic to the level of the toll.

2.3 Experiment

The experiment was conducted at the CREED laboratory at the University of Amsterdam in February 2012 with a total of 240 subjects recruited from the CREED subject pool. Participants were mainly undergraduate students, slightly less than half were female and roughly 60% majored in economics or business. The experiment was programmed in PHP/MySQL. Every participant received a show up fee of 7 euros. During the experiment, ‘points’ were used as currency. These were exchanged into euros at the end of each session at an exchange rate of 1 euro per 600 points. The experiment lasted between 90 minutes and 2 hours and participants earned on average about 22 euros, including the show up fee. Before starting, the participants had to answer control questions to make sure that they understood the instructions. The experiment did not start until all participants had successfully answered these questions. Appendix 2.A provides a transcript of the instructions and test questions, Appendix 2.B contains screenshots. During the experiment, subjects received no information on the choices or the performance of other subjects. Twelve sessions were run, three each for four distinct treatments.¹¹ When scheduling, we distributed the treatments in a balanced way over mornings and afternoons and across the different days of the week.

2.3.1 Treatments

The design is a 2×2 factorial, between-subject design. Subjects are either employer or employee. They are allocated to groups consisting of one employer and five employees. Those in the role of employees work on a task for which they receive performance-based remuneration; the employer receives earnings depending on the performance of the employees in

¹¹A pilot was run in the summer of 2011, as documented in Weber (2011). The new sessions differ significantly from the pilot; the most important changes are the introduction of a mechanism to measure preferences for the size of the public sector and the introduction of a leisure task. More information is available upon request.

the same group. The form this incentive scheme takes is one of the treatment variables. In one case, employees receive a gross wage, from which a duty is subsequently deducted as a tax. In the other, employers pay the duty and employees receive a (lower) net wage. Note that this corresponds naturally to labor market taxes levied either on the employees' or on the employers' side. What happens with the tax proceeds is varied in the second treatment variable. The tax proceeds are either taken away ('nothing in return') or used to produce a public good (which is called 'common fund' in the experiment). Table 2.1 summarizes the design. The acronyms *EN*, *IN*, *EP* and *IP* for the four distinct treatments will be used regularly in the remainder.

Table 2.1: 2×2 design

	Employer payroll tax	Income tax
Nothing in return	<i>EN</i> (60)	<i>IN</i> (60)
Public good	<i>EP</i> (60)	<i>IP</i> (60)

Notes: Cells indicate the acronyms used for the treatment. Each treatment combines a tax levied either on the employers' (*E*) or employees' (*I*) side with the case where tax proceeds are either lost (*N*) or used to produce a public good (*P*). In parentheses are the numbers of subjects per treatment.

The tax rate used is 40% in the income tax treatments and 66.7% of the corresponding lower wage in the employer payroll tax treatments. The public good in the relevant treatments is produced with a multiplication factor of 1.3 (meaning that the tax revenue allocated to the common fund is increased by 30%) and its returns are equally distributed among all employees in a group at the end of the experiment.¹² The wages were chosen such, that the net wage in a nothing-in-return treatment is equal to the net wage plus the return from the public good from one's own tax payment in the public good treatments.¹³ As we will explain

¹²As a consequence, a tax revenue of r points in a group with 5 employees yields $\frac{1.3r}{5}$ points from the public good for each employee. The public good is not supplied to employers, because we envisage a public good related to income security, e.g., unemployment benefits. Though the public good treatments involve (mandatory) contributions, this experiment is ill-suited to isolate free-riding motives because many factors are involved in the decision to work (and thus contribute via taxes to the public good). Our results will show that labor supply and job performance are not higher in the public good treatments than in the corresponding nothing in return treatments which suggests that subjects are not focusing on payments to others from the public good when they decide about labor supply and effort. Indirectly, this could be interpreted as some evidence for free riding.

¹³Recall that in the public good treatments the net earnings consist of the net wage plus the returns from the public good. The returns from the public good can be split in a part that is due to own tax payments and a part that is due to the taxes paid by the other employees in the group. Consider a task an employee faces. Denote now by W_P and W_N the net wages for this task in the two public good and the two nothing in return treatments, respectively. Denote by $\gamma \cdot W_P$ the tax paid, which is a (mandatory) contribution to the public good in the public good treatments (γ thus corresponds to the tax rate in the employer payroll tax treatments but not in the income tax treatments, because there the tax base is not the net wage). In the nothing in return treatments the employee receives W_N , which compares to $W_P + \frac{\gamma \cdot W_P \cdot 1.3}{5}$ in the public good treatments. W_P and W_N are chosen to equate these two returns.

below, the payment schedule for correctly solved problems of the real effort work task is decreasing. In the public good treatments subjects receive (untaxed) additional earnings from the public good (i.e. from the performance by others) after the experiment, but they do not receive any information on others' performances while taking their decisions. To illustrate tax equivalence between treatments, Table 2.2 shows an example of wages, taxes, and net earnings in the experiment. In this example it is assumed that one employee solves only the first problem correctly (the tax schemes are equivalent as illustrated, no matter how many problems are solved correctly).

Table 2.2: Illustration of wages, taxes, and earnings in the experiment

	<i>EN</i>	<i>EN</i>	<i>EP</i>	<i>IP</i>
Gross wage	280.8	468	239.32	398.87
Income tax (40%)		-187.2		-159.55
Net wage	280.8	280.8	239.32	239.32
Own performance PG benefits			41.48	41.48
Net earnings employee	280.8	280.8	280.8	280.8
Employer tax (66.7% of gross wage)	-187.2		-159.55	
Total labor costs (wage + tax)	468	468	398.87	398.87
Net earnings employer	49.8	49.8	49.8	49.8

Notes: This table illustrates the equivalence of the taxes in the different treatments for the example that an employee has solved the first problem correctly. The numbers are in points.

2.3.2 Course of Events

At the beginning of each session, subjects are randomly divided into groups of six. One subject in each group is randomly determined to be 'employer', the other five are 'employees'. The group composition remains fixed throughout the experiment. The experiment consists of multiple parts. The participants receive the instructions for a part only after the previous parts have been completed. The terms 'employer' and 'employee' are used intentionally, as is the term 'wage'. Neutral wording is chosen for the duties in order to avoid (unmeasured) preconceptions that some subjects might have with respect to terms referring directly to taxes. Figure 2.1 gives a schematic overview of the timeline; the different parts will be explained in the following.

Real-Effort Task and Leisure Option

The experiment involves a real-effort work task, which is the following. Each employee sees two 10×10 matrices on the screen that are filled with randomly generated two-digit

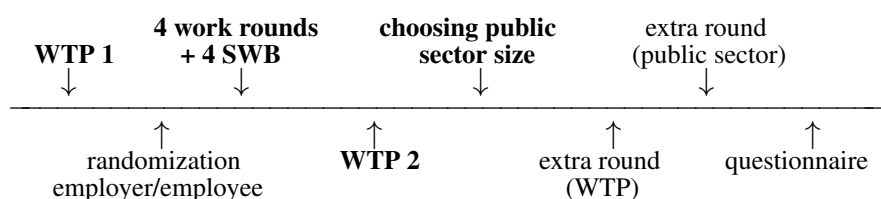


Figure 2.1: Timeline of the experiment

Notes: The parts encompassing decisions or actions leading to dependent variables are in bold. The acronyms depict the elicitation of subjects' willingness to pay for an extra round (WTP) and subjective well-being (SWB). Instructions are only given right before the respective task, except for the beginning where instructions for WTP 1 and the four regular rounds are given. Subjects receive information on the randomly drawn price for the first and the parameters for the second extra round right before the respective round starts. Information on (payments due to) other subjects' decisions is only provided right before the questionnaire.

numbers. Figure 2.2 shows a screenshot from the work task (taken from treatment *IN*; see Appendix 2.B for a larger version). The employees' job is to find the largest number in the left matrix and the largest number in the right matrix and to add these two numbers up. For the summation, the participants are provided with pocket calculators. After answering, irrespective of whether the answer is correct, a new pair of matrices appears. This means that subjects have only one attempt to provide the correct answer. Each employee faces a maximum of 30 of these problems, which is much more than they can actually solve correctly in one round, which lasts for 8 minutes. This limit and the way the random numbers are generated make guessing a very unsuccessful strategy. Only the number of correct additions matters, there is no punishment for incorrect additions. While the employees are doing this task, they can see at the top of the monitor the amount they will receive if the next number they enter is correct and, where applicable (i.e., when tax is levied on the employees), the amount that will be deducted from it (as a tax). Furthermore, they can see how much they have already earned and, where applicable, the amount that will be deducted from it. They can also see the number of correct and incorrect additions so far and the remaining time. The total number of correctly solved problems is a measure of job performance. This procedure is repeated in four independent and identical rounds. During these rounds, employers do not need to do anything. Employers are passive in the experiment, but they nevertheless play an important role for two reasons. First, their presence increases the external validity of the framing in terms of wages and tasks. Second, subjects with other-regarding preferences are affected differently when (as in the world outside of the laboratory) their decisions have real consequences for another person (Charness et al., 2007; Sutter, 2009).

Employers receive a net payment of 49.8 points per correct addition by any of their five employees. Net payments for the employees are linearly decreasing in the number of

Amount of correct additions in this round so far:	1	Round 1 of 4	Time: 06:53
Amount of incorrect additions in this round so far:	1		
Remaining problems that can be solved in this round:	28	Total wage earned in this round:	429 (minus 171.6)

If your next answer is correct you will receive 390 points (from which 156 will be deducted).

<table border="1"> <tr><td>38</td><td>39</td><td>24</td><td>30</td><td>58</td><td>67</td><td>53</td><td>60</td><td>53</td><td>58</td></tr> <tr><td>73</td><td>20</td><td>12</td><td>79</td><td>56</td><td>32</td><td>29</td><td>24</td><td>79</td><td>40</td></tr> <tr><td>24</td><td>30</td><td>43</td><td>25</td><td>21</td><td>59</td><td>25</td><td>73</td><td>78</td><td>41</td></tr> <tr><td>67</td><td>23</td><td>10</td><td>77</td><td>35</td><td>30</td><td>15</td><td>25</td><td>40</td><td>76</td></tr> <tr><td>48</td><td>58</td><td>20</td><td>36</td><td>28</td><td>35</td><td>15</td><td>29</td><td>49</td><td>57</td></tr> <tr><td>33</td><td>71</td><td>30</td><td>53</td><td>44</td><td>72</td><td>65</td><td>55</td><td>56</td><td>49</td></tr> <tr><td>73</td><td>62</td><td>39</td><td>50</td><td>78</td><td>35</td><td>72</td><td>22</td><td>69</td><td>44</td></tr> <tr><td>54</td><td>14</td><td>71</td><td>50</td><td>63</td><td>42</td><td>71</td><td>52</td><td>70</td><td>17</td></tr> <tr><td>25</td><td>12</td><td>13</td><td>23</td><td>62</td><td>68</td><td>71</td><td>41</td><td>65</td><td>43</td></tr> <tr><td>42</td><td>16</td><td>24</td><td>74</td><td>38</td><td>68</td><td>32</td><td>56</td><td>65</td><td>74</td></tr> </table>	38	39	24	30	58	67	53	60	53	58	73	20	12	79	56	32	29	24	79	40	24	30	43	25	21	59	25	73	78	41	67	23	10	77	35	30	15	25	40	76	48	58	20	36	28	35	15	29	49	57	33	71	30	53	44	72	65	55	56	49	73	62	39	50	78	35	72	22	69	44	54	14	71	50	63	42	71	52	70	17	25	12	13	23	62	68	71	41	65	43	42	16	24	74	38	68	32	56	65	74	<table border="1"> <tr><td>83</td><td>57</td><td>25</td><td>13</td><td>51</td><td>34</td><td>66</td><td>42</td><td>74</td><td>63</td></tr> <tr><td>12</td><td>69</td><td>29</td><td>28</td><td>19</td><td>37</td><td>53</td><td>77</td><td>57</td><td>74</td></tr> <tr><td>22</td><td>22</td><td>58</td><td>46</td><td>75</td><td>82</td><td>66</td><td>82</td><td>13</td><td>70</td></tr> <tr><td>14</td><td>50</td><td>47</td><td>58</td><td>21</td><td>76</td><td>56</td><td>81</td><td>19</td><td>30</td></tr> <tr><td>27</td><td>62</td><td>52</td><td>38</td><td>59</td><td>36</td><td>54</td><td>42</td><td>68</td><td>74</td></tr> <tr><td>67</td><td>21</td><td>78</td><td>64</td><td>25</td><td>67</td><td>78</td><td>77</td><td>46</td><td>83</td></tr> <tr><td>20</td><td>73</td><td>63</td><td>13</td><td>40</td><td>51</td><td>45</td><td>57</td><td>15</td><td>45</td></tr> <tr><td>74</td><td>83</td><td>81</td><td>52</td><td>46</td><td>71</td><td>43</td><td>34</td><td>42</td><td>37</td></tr> <tr><td>37</td><td>62</td><td>31</td><td>34</td><td>13</td><td>23</td><td>20</td><td>78</td><td>36</td><td>62</td></tr> <tr><td>54</td><td>75</td><td>55</td><td>67</td><td>65</td><td>40</td><td>59</td><td>34</td><td>23</td><td>28</td></tr> </table>	83	57	25	13	51	34	66	42	74	63	12	69	29	28	19	37	53	77	57	74	22	22	58	46	75	82	66	82	13	70	14	50	47	58	21	76	56	81	19	30	27	62	52	38	59	36	54	42	68	74	67	21	78	64	25	67	78	77	46	83	20	73	63	13	40	51	45	57	15	45	74	83	81	52	46	71	43	34	42	37	37	62	31	34	13	23	20	78	36	62	54	75	55	67	65	40	59	34	23	28
38	39	24	30	58	67	53	60	53	58																																																																																																																																																																																																
73	20	12	79	56	32	29	24	79	40																																																																																																																																																																																																
24	30	43	25	21	59	25	73	78	41																																																																																																																																																																																																
67	23	10	77	35	30	15	25	40	76																																																																																																																																																																																																
48	58	20	36	28	35	15	29	49	57																																																																																																																																																																																																
33	71	30	53	44	72	65	55	56	49																																																																																																																																																																																																
73	62	39	50	78	35	72	22	69	44																																																																																																																																																																																																
54	14	71	50	63	42	71	52	70	17																																																																																																																																																																																																
25	12	13	23	62	68	71	41	65	43																																																																																																																																																																																																
42	16	24	74	38	68	32	56	65	74																																																																																																																																																																																																
83	57	25	13	51	34	66	42	74	63																																																																																																																																																																																																
12	69	29	28	19	37	53	77	57	74																																																																																																																																																																																																
22	22	58	46	75	82	66	82	13	70																																																																																																																																																																																																
14	50	47	58	21	76	56	81	19	30																																																																																																																																																																																																
27	62	52	38	59	36	54	42	68	74																																																																																																																																																																																																
67	21	78	64	25	67	78	77	46	83																																																																																																																																																																																																
20	73	63	13	40	51	45	57	15	45																																																																																																																																																																																																
74	83	81	52	46	71	43	34	42	37																																																																																																																																																																																																
37	62	31	34	13	23	20	78	36	62																																																																																																																																																																																																
54	75	55	67	65	40	59	34	23	28																																																																																																																																																																																																

Largest number in the left matrix plus largest number in the right matrix:

Figure 2.2: Screenshot during a work round (taken in treatment *IN*)

attempts (but are restricted to be non-negative). If employees solve the first problem correctly in the nothing in return treatments, they receive 280.8 points (cf. Table 2.2). With each attempt (whether correct or not) the payment for the next correct addition decreases by 23.4 points. In the public good treatments, these numbers correspond to the net return from own performance (the direct net wage plus the return from the public good that is due to own performance).¹⁴

We provide employees with an outside option. Instead of working, they can also choose

¹⁴Decreasing payments make it likely that subjects will use the fixed payment option (the leisure task) at some point, i.e. they lead to a large number of interior solutions concerning the time spent working. They can be seen as representing diminishing marginal revenue. Formally, net earnings from correctly solving a problem in the nothing in return treatments can be written as $\pi = \max(280.8 - 23.4x, 0)$, where x is the number of problems the employee has previously attempted to solve in the same round. This is also the gross wage in *EN* (net wage equals gross wage), while the gross wage in the *IN* is $\pi = \max(468 - 39x, 0)$, which leads, with a tax rate of 40%, to the same net wage as in *EN*.

In the public good treatments subjects receive additional earnings from the tax payments by others (which they cannot influence in any way). Over the four regular work rounds these extra earnings lie between 0 (if no other group member solves any problem correctly) and 4314 points (if all co-workers solve all problems correctly).

a leisure task, which is framed as a ‘fixed payment option’. At any moment during the work rounds, employees can click on a button ‘Go to fixed payment option’. After doing so, they are shown a largely empty screen for the rest of the round and receive a fixed payment of 2.2 points per second remaining. They cannot return to solving problems in the same round. Note that the total amount of time (in seconds) spent in the ‘work-mode’ provides a natural measure of labor supply at the intensive margin.

Measuring Subjective Well-Being

After each round, the employees are shown a screen depicting their gross wage and the number of points paid as tax (if applicable) in the preceding round. Then participants are surveyed to measure their subjective well-being using a self-assessment manikin (the SAM-V-9; Irtel, 2007, Lang, 1985, Bradley and Lang, 1994). This measure of subjective well-being is also referred to in the literature as satisfaction, happiness, or experienced pleasure. Subjects are asked to report how they are feeling by clicking on one of nine images on the manikin. These images are drawings depicting emotions ordered from least happy to most happy, thus yielding a score from 1 (low pleasure) to 9 (high pleasure). The number is referred to as the ‘self-assessment score’. We will use the sum of these scores over the four rounds as our measure of subjective well-being. The self-assessment manikin is shown in Figure 2.3.

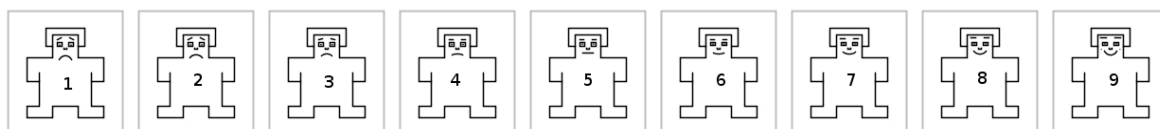


Figure 2.3: Subjective well-being self-assessment manikin

Notes: After each round, subjects are asked to choose the one of the nine figures that best describes their current emotion.

Measuring Labor Supply at the Extensive Margin

After finishing the instructions for the part comprising the four work rounds, but before being told whether they are employers or employees, subjects are asked to state their willingness to pay for participating in an extra work round after the four regular rounds will have been completed.¹⁵ For this purpose a BDM mechanism is used (Becker et al., 1964). The price of participation in the extra round is determined randomly (drawn from a uniform distribution

¹⁵For subjects subsequently randomized to be employers the revealed willingness to pay has no further consequence.

between 1056 and 2400 points). The lower limit corresponds to the amount earned after immediately choosing the fixed payment option, the upper limit is a number slightly higher than our expectations of maximum possible earnings in one round. It is randomly determined whether the extra round takes place or not. If it takes place and if the price is lower than the amount stated by an employee, this employee pays the price and works (and gets paid) for another round. If the price is higher than the bid of the employee, the employee neither pays for nor works in another round. Subjects not participating remain seated until all participants have finished. If subjects have a true valuation for participating in this extra round it is a dominant strategy to bid this true valuation. In the treatments with public good the returns from the public good are split among all employees of a group, those working in the extra round and those not working. The stated willingness to pay provides us with a (first) measure of labor supply at the extensive margin.¹⁶

After having finished the regular four work rounds, employees are confronted with the same BDM mechanism again. They are told that the extra round corresponding to the number they enter after having completed the regular rounds will be played out if and only if the extra round corresponding to the willingness to pay elicitation before the regular rounds will not be played out. This willingness to pay after the four work rounds provides us with a (second) measure of labor supply at the extensive margin. Our procedures imply that an extra round based on stated willingness to pay always takes place, at a later point (though, whether an individual employee participates in it depends on her stated willingness to pay). It consists of exactly one round, based on either the first (pre-play) BDM mechanism or the second (post-play).

Eliciting Preferences for Public Sector Size

After this second statement of willingness to pay, subjects are told that there will be yet another round. All employees participate in this round, which will take place with new rules. The rules differ from the regular rounds as follows. In the nothing in return treatments (*EN* and *IN*), a public good is introduced, such that the taxes are no longer lost, just as in the public good treatments. Now, the tax rate and the multiplication factor for the public good are no longer given. Instead, they are chosen by the employees in a random dictator style, using the following mechanism. Subjects are presented with a slider, as shown in Figure 2.4. Each position of the slider represents a unique combination of tax rate and

¹⁶In reality, people usually do not start participating in the labor market by just clicking on a button somewhere. There are all kinds of costs associated with beginning a new job, including looking for suitable job offers, writing and sending applications, going to interviews, maybe even moving to a different city, etc. Therefore, we consider this willingness to pay a better measure of labor supply at the extensive margin than a mere statement whether one wants to participate or not.

multiplication factor. At the left end of the slider, the tax rate is zero, while the multiplication factor for the public good is high. When moving the slider from left to right, the tax rate increases, and the multiplication factor decreases. The increase in the tax rate is different for the income tax and employer payroll tax treatments, because the tax bases are different – the tax rates are chosen in a way that each position of the slider leads to the same tax revenue (as a percentage of labor costs; i.e. to the same absolute tax revenue in the absence of behavioral responses to the tax framing). The trade-off between tax rate and multiplication factor can be interpreted as a diminishing marginal productivity of the public sector; the higher the tax revenue is, the lower is the efficiency of public good production. While one could also have subjects choose the tax rate for a fixed multiplication factor, we prefer to explicitly allow for this trade-off in order to accommodate preferences for a point at the interior of the slider. The slider has 101 different positions, yielding a number between 0 and 100, where 0 corresponds to the leftmost position of the slider, which is used as the default position.¹⁷ The number corresponding to the chosen slider position provides us with a measure of the subject's preferred size of the public sector. After all subjects have chosen a slider position, one employee in each group is randomly selected and her choice is used for this extra round. Note that the employer payroll tax and the income tax treatments are still absolutely equivalent, whereas the nothing in return and the public good treatments are now somewhat 'less equivalent'. Subjects in the nothing in return treatments have had no experience with the public good prior to this round. Furthermore, net payments in the nothing in return treatment are not adjusted to the levels used in the public good treatment in order to avoid subjects having to adapt to a new payment schedule for the extra round. As a consequence, payoffs here are slightly higher in the treatments *EN* and *IN* than in *EP* and *IP*.

Finally, the two extra rounds are played out (one originating from the willingness to pay for an extra round and one from the choice of public sector size parameters).¹⁸ The information on the randomly drawn price and the public sector size parameters selected in

¹⁷The multiplication factor of the public good is 3 at the default position on the left end and 0.75 at the right end. The tax is 0 at the left end and corresponds to 75% of the employer's labor cost at the right end; the tax is always expressed as a percentage of the employee's (gross) wage. Note that choosing a position where the multiplication factor is less than 1 is dominated in the sense that subjects would always earn more at lower tax/higher productivity rates.

If there is an anchoring effect where subjects choose the tax rates that they experienced in the earlier work rounds (40% in the income tax treatments and 66.7% (of a lower tax base) in the employer payroll tax treatments), this would lead to the same slider position in all treatments.

¹⁸Before these rounds are played, we elicit loss aversion using the test developed by Fehr and Goette (2007); see Appendix 2.A. The measure turns out not to lead to additional insights in the data analysis – we attribute this to the fact that we are not dealing with risky choices (for which the test was designed) and to the fact that the measure is quite rough (most subjects end up in the same category). Therefore, this measure will not be discussed further in this chapter. More information is available from the author.

Please choose your most preferred position of the slider.

Multiplication factor of the common fund:	3.00
Contribution to the common fund, expressed as a percentage of your wage:	0.00




Figure 2.4: Slider for the elicitation of public sector size preference

Notes: The slider is at the default position. When moving the slider from left to right the multiplication factor of the public good (framed as a ‘common fund’) decreases while the contribution to the public good (i.e. the tax rate) increases.

random dictator style is given to subjects before the respective round starts. After these two rounds information on payments stemming from others’ public goods contributions (i.e., their tax payments) is revealed. Before being paid, participants are asked to fill out a short questionnaire, including questions concerning gender, age, field of study and experience in laboratory experiments (see Appendix 2.A).

2.4 Hypotheses and Mechanisms

The main goal of this chapter is to test the effects of the liability side of a labor tax – represented by pure framing – on outcomes of the three dimensions we discussed. We use null-hypotheses about treatment effects arising from rational choice behavior. These imply no differences across treatments.¹⁹ However, note that the two taxes are also equivalent assuming various other outcome based preferences, such as for example inequity aversion (Fehr and Schmidt, 1999). The incentive schemes we use are equivalent by design, which means that no general equilibrium mechanisms are needed to arrive at the full-rationality outcome.

¹⁹As we are mainly interested in the effects of the tax liability side, we do not formulate or test hypotheses between the nothing in return and the public good treatments. While the treatments are all equivalent with selfish and fully rational agents, there can be differences between a nothing in return and a corresponding public good treatment with boundedly rational agents (for example if a dollar of wage is more salient than a dollar received from the public good due to own performance).

Strictly speaking, there are no rational choice predictions for subjective well-being, because decisions are not incentivized. We also use the null-hypotheses of no difference in outcomes here.

Now, we discuss which psychological mechanisms we expect may govern participants' behavior. These mechanisms arise partly from salience considerations (Chetty et al., 2009; Finkelstein, 2009; Fochmann and Weimann, 2013; Rupert and Wright, 1998; Sausgruber and Tyran, 2005, 2011), but they are not exclusively based on tax salience. The mechanisms are the following:

- (i) We regard a dollar of wage more salient than a dollar of tax. As a consequence, people will tend to focus more on gross wages than on net wages (as they do not fully take the taxes into account), which is what Fochmann et al. (2010b) call 'gross wage illusion' (as observed by Fochmann et al., 2010b, and Hayashi et al., 2013, in different settings).
- (ii) We expect people to consider a tax payment as a loss that, *ceteris paribus*, they would prefer to avoid. This effect is expected to be asymmetric, in that an employee sees a tax paid by herself as more of a loss than a tax paid by her employer. We call this 'tax loss effect'.
- (iii) Individuals having other-regarding preferences may derive positive utility when a public good is provided to others using tax payments they made ('warm glow', Andreoni, 1990). This warm glow effect here depends on the perception of the tax – it exists if individuals feel more that it is actually their money that is used to benefit others under an income tax than under an employer payroll tax.

Each of these mechanisms has different implications under an employer payroll tax than under an income tax. Before we discuss in more detail how these mechanisms interact and influence participants' behavior in the decisions and tasks we are investigating, we first briefly describe the consequences of each mechanism in more general terms. Gross wage illusion leads individuals to overestimate their net earnings and therefore makes them value work (including payment) higher under an income tax than under a payroll tax. The tax loss effect makes individuals value work (including payment) less under an income tax than under an employer payroll tax, because they dislike the fact that they have to bear the taxes. Warm glow only exists when taxes are used as contributions to the production of public goods and leads to individuals valuing work (including payment) more when the tax is levied on the employee's side, as they then have the feeling of doing a good deed. We allow each of these mechanisms to have an impact that differs across the decisions and tasks participants face.

We start with preferences regarding the size of the public sector. Gross wage illusion may lead to a higher preferred size in the income tax treatments than in the payroll tax treatments. This is the case if there is a positive income effect for the demand for public goods, i.e., higher (perceived) income yields a preference for a larger public sector (although possible, we expect this effect to be very small if it exists).²⁰ The tax loss effect has the opposite effect,

²⁰The normality of public goods is a common assumption (e.g., Sugden, 1984, Morton, 1987, or Andreoni,

however, because it implies that for any given size of the public sector the perceived tax costs needed to achieve it are smaller when taxes are levied on the employers' side of the labor market. Hence, a larger public sector is preferred in the payroll tax treatments. The warm glow effect goes in the same direction as gross wage illusion. Individuals subject to warm glow may prefer a larger public sector when they are funding it. With two effects pointing in the direction of a higher preferred size in the income tax treatment and one effect predicting the opposite, the aggregate effect could go either way.²¹

Next, consider subjective well-being. In the nothing in return treatments, gross wage illusion implies that people perceive a higher income in the income tax treatment than in the payroll tax treatments. A natural hypothesis is then that they will feel better in the former case. The tax loss effect points in the opposite direction: people will be less happy when they feel that the tax represents money taken away from them. If there is a public good, also warm glow may be present. This effect aligns with the effect of gross wage illusion on subjective well-being. It predicts that people will feel better in the income tax case, because they conceive themselves as helping.

Finally, for labor supply and job performance, we again need to consider the aggregate of the three effects. Gross wage illusion directly implies that people will supply more labor (at both the intensive and extensive margins) and perform better in the income tax treatments.²² The tax loss effect affects labor supply decisions in the opposite direction. This is because such decisions are assumed to be based on the (subjective) costs and benefits of exerting effort. Taxes on the employees' side add to the perceived costs of exerting effort more than taxes on the employers' side. In the treatments with a public good the warm glow effect will yield higher labor supply and better performance in the income tax treatment because of the perceived extra benefits related to helping others. Once again, gross wage illusion and warm glow point in one direction, and the tax loss effect points to the opposite.

On top of the fact that we expect these mechanisms to lead to the rejection of rational behavior null-hypotheses, one part of these mechanisms can be tested directly within the framework of this experiment. The warm glow effect when considering subjective well being or labor supply and job performance only exists in the public good treatments. Thus, if our mechanisms drive the results and warm glow plays a role, there should be a positive interaction effect between the income tax treatment condition and the public good treatment condition for subjective well-being, labor supply, and job performance. We will test this,

1990). For indirect empirical support, see, e.g., Schram (1990).

²¹Note that we assume that the framing of the tax has no influence on subjects' beliefs on how others react to the tax.

²²This is the case if a higher wage leads to higher labor supply, as seems to be generally the case in the real world (Evers et al., 2008) and arguably even more so in laboratory experiments.

below.²³

2.5 Results

We have collected data from 10 groups per treatment. Because we only use data obtained from employees to test our hypotheses, this gives 50 observations per treatment, except for the first measure of labor supply at the extensive margin, where we have 60 observations per treatment (as this was measured before the randomization into employees and employers).²⁴ Because subjects receive no feedback about others' decisions until the end of the experiment, we can treat observations across individuals as statistically independent. The outcome variables we consider are preferences for the size of the public sector (as measured by the chosen slider position), subjective well-being (as measured by the sum of the choices with the self-assessment manikin), both measures of labor supply at the extensive margin (as given by choices in the two BDM tasks), labor supply at the intensive margin (as measured by the number of seconds in the work mode over the four regular rounds), and job performance (as measured by the total number of problems solved correctly).

We present both non-parametric tests and censored regression models in the analysis of our results. Regressions allow us to observe the effects of the interaction between the tax type and the use of the tax proceeds and to control for observable characteristics such as age and gender.²⁵ As is common in experimental papers, we focus on the directions of treatment effects rather than on levels and quantifications (the meaning of exact quantities in laboratory experiments is often limited).

2.5.1 Public Sector Size Preference

Table 2.3 shows descriptive statistics and p-values from two-sided Wilcoxon rank sum tests. Table 2.4 shows the coefficient estimates from censored regressions (both with and without covariates).²⁶

²³Note that our mechanisms do not predict an interaction effect between income tax and public good treatment for public sector size preference. This is because in the corresponding extra round a public good is also introduced in treatments *EN* and *IN*. Without a public good it would always be optimal to chose a zero tax rate.

²⁴Due to computer problems, one observation in *IN* has missing data on subjective well-being and one observation in *IP* has missing data on subjective well-being, intensive labor supply, and job performance.

²⁵Appendix 2.C.1 also presents the data split according to gender. We observe gender effects for some of our results. This could mean that the mechanisms we distinguished between might affect men differently than women. We did not intend to study gender effects and therefore the analysis remains explorative, however.

²⁶The optimal slider position cannot easily be determined in the experiment as it depends on subjects' reactions to the tax (and to the use of the tax proceeds and to the tax framing). If labor supply/performance were completely inelastic to the tax the optimal slider position would be 45 (which corresponds to an income

Table 2.3: Overview and tests, public sector size preference

	Employer payroll tax mean (std. error)	Income tax mean (std. error)	Treatment diff. p-value Wilcoxon
Nothing in return	51.18 (4.22)	33.00 (2.77)	0.002
Public good	41.18 (3.69)	33.12 (2.96)	0.080

Notes: Individual outcomes are integers between 0 and 100, larger numbers representing preference for a larger public sector. The p-values stem from two-sided Wilcoxon rank sum tests.

Both, when taxes are lost and when proceeds are invested in a public good, subjects prefer a larger public sector when the tax is levied on the employer's side. These differences are highly significant in the case where tax proceeds are lost and marginally significant between the public good treatments. The results from the regression analysis strengthen this result – the coefficient of the tax condition dummy is negative (and significant at the 1%-level, independently of whether or not covariates are included in the regression). The interaction coefficient is not significant, which is not particularly surprising, given that the mechanisms we outlined in Section 2.4 do not predict an interaction effect for this dependent variable. There are (marginally) significant negative effects of age and being in the public good treatment (the latter is even significant at the 5%-level if no covariates are included in the regression). This dummy measures whether or not the subject participated in the public good treatment before this preference was measured. The result implies that having experienced the public good leads one to prefer a smaller public sector.

Recalling that of the three mechanisms we discussed, gross wage illusion and warm glow both predict that a larger public sector is preferred in the income tax treatment, our findings suggest that the tax loss effect plays a dominant role when individuals determine their preference for the size of the public sector.

2.5.2 Subjective Well-Being

Table 2.5 gives an overview of the results on subjective well-being. Table 2.6 shows the corresponding regression results.²⁷

We can see in Table 2.5 that participants report higher subjective well-being under an em-

tax of 33.75 percent). Given that a substantial part of the tax proceeds is paid back to an employee via the public good (around 40 percent of the paid tax at the position 45, more if it is below) and given the decreasing marginal pay for completing the work task, it may be that a true optimum in the experiment is not far below 45. It is important, however, to keep in mind that relative positions with respect to such an optimum do not arise from treatment effects. Because the comparative statics of treatment comparisons are generally considered to be most relevant for behavior outside of the laboratory, it is questionable whether the relative position to the optimum has much meaning for the world outside of the laboratory.

²⁷Consideration of the four separate self-reports shows very consistent behavior and no indication of trends.

Table 2.4: Regression results for public sector size preference

	Coefficient	Std. error	Coefficient	Std. error
Tax (1=income tax)	-20.10***	(5.48)	-17.15***	(5.54)
PG (1=public good)	-11.50**	(5.49)	-9.91*	(5.46)
Tax * PG	11.62	(7.73)	9.04	(7.77)
Intercept	52.31***	(3.89)	72.30***	(19.58)
Age			-1.40*	(0.83)
Gender (1=male)			4.18	(4.03)
Studies (1=econ+science)			-2.41	(4.50)
Lab experience (1=yes)			9.67	(6.04)
Observations	200		200	

Notes: The table shows outcomes of a Tobit regression. Individual outcomes are integers between 0 and 100. 13 observations are left censored and 11 right censored. *, **, and *** stand for significance at the 10-, 5-, and 1-percent level.

Table 2.5: Overview and tests, subjective well-being

	Employer payroll tax mean (std. error)	Income tax mean (std. error)	Treatment diff. p-value Wilcoxon
Nothing in return	21.84 (0.66)	18.94 (0.84)	0.007
Public good	21.16 (0.77)	21.82 (0.75)	0.594

Notes: Individual outcomes are integers between 4 and 36, larger numbers representing higher subjective well-being. The p-values stem from two-sided Wilcoxon rank sum tests.

ployer payroll tax when the tax proceeds are lost. This finding is significant at the 1%-level. Subjective well-being is not affected by the tax treatment when the tax proceeds are invested in a public good. In line with the results from the non-parametric tests, the regression results show a highly significant tax dummy coefficient. Furthermore, the interaction effect between the tax treatment dummy and the public good dummy is positive and significantly different from zero (at the 5%-level). This shows that the effect of larger subjective well-being under an employer payroll tax is weakened or even reversed when tax proceeds are used beneficially (for our parameters, the two effects are of more or less equal magnitude; this explains the outcomes in Table 2.5).

Looking back at our mechanisms, these results suggest that when tax proceeds yield nothing in return, the tax loss effect outweighs the effect of gross wage illusion on subjective well-being (because the coefficient of the tax dummy is significantly negative). Warm glow only plays a role when the tax proceeds are contributed to a public good where it predicts higher subjective-well being under an income tax (thus a positive interaction term). The finding of the interaction effect thus corroborates our view that the mechanisms we described

Table 2.6: Regression results for subjective well-being

	Coefficient	Std. error	Coefficient	Std. error
Tax (1=income tax)	-2.90***	(1.06)	-2.80***	(1.08)
PG (1=public good)	-0.71	1.06	-0.62	(1.05)
Tax * PG	3.59**	(1.50)	3.09**	(1.51)
Intercept	21.84***	(0.75)	23.81***	(3.80)
Age			-0.04	(0.16)
Gender (1=male)			1.54**	(0.78)
Studies (1=econ+science)			-0.61	(0.87)
Lab experience (1=yes)			-1.79	(1.17)
Observations		198		198

Notes: The table shows outcomes of a Tobit regression. Individual outcomes are integers between 4 and 36. 1 observation is left censored, none are right censored. *, **, and *** stand for significance at the 10-, 5-, and 1-percent level.

are at work.

2.5.3 Labor Supply and Job Performance

The measures on labor supply and job performance we use can be divided into two parts. The first consists of two measures of labor supply at the extensive margin (once extracted before the randomization and experiencing the work task and once after). These measures capture the willingness to bear a cost in order to work. The extensive labor supply decisions are taken outside of the regular work environment in our experiment, but arguably also in the world outside of the laboratory.²⁸

The second part of our measures concern labor supply at the intensive margin and job performance. These correspond to decisions/performance made while working. In our experiment, subjects are solving problems while time is running down and while we are measuring how well they do the task. If they get annoyed, stressed, or bored, or if they no longer deem the task profitable, they can stop supplying labor by quitting the task for the remainder of the current round. Labor supply at the intensive margin and job performance are of course related; participants spending more time solving problems are likely to solve more of them. They also differ substantially from the two measures of labor supply at the extensive margin. For this reason, we will separate the discussion and first discuss labor supply at the extensive

²⁸Outside of the laboratory, this decision will often be made when people know which kinds of job they are applying to, but do not really know what they can precisely expect. This corresponds to our first measure. Our second measure is similar, but measured after four rounds of work (in the outside world, this corresponds to the decision to apply to a similar job, if available, after termination of a job).

margin and then labor supply at the intensive margin and job performance.²⁹

Labor Supply at the Extensive Margin

Table 2.7 gives an overview for the outcomes on labor supply at the extensive margin. It shows that independently of which measure one considers and of whether one considers the nothing in return treatments or the public good treatments, labor supply is always higher under an income tax. However, while the differences are always considerable (more than two standard errors in two cases and still more than one standard error in the other two cases), only one of the differences is statistically significant at the 5%-level when tested with a Wilcoxon rank sum test. Considering Table 2.7, there seems thus to be some (admittedly rather weak) evidence that extensive labor supply is higher under an income tax than under an employer payroll tax.

Table 2.7: Overview and tests, labor supply at the extensive margin

	Employer payroll tax mean (std. error)	Income tax mean (std. error)	Treatment diff. p-value Wilcoxon
Measure 1			
Nothing in return	1416.3 (34.8)	1525.4 (46.1)	0.133
Public good	1408.8 (39.8)	1462.7 (40.8)	0.268
Measure 2			
Nothing in return	1273.9 (45.3)	1347.0 (49.7)	0.167
Public good	1229.2 (39.2)	1328.0 (38.6)	0.035

Notes: Individual outcomes are integers between 1056 and 2400, larger numbers represent higher labor supply. The p-values stem from two-sided Wilcoxon rank sum tests.

Tables 2.8 and 2.9 show the regression results, separately for both measures. The regression results for the first measure show a positive coefficient of the tax dummy that is significantly different from zero (only marginally so if no covariates are included in the regression). Furthermore, looking at the regression coefficients, one can see that there is a strong and statistically highly significant positive effect of being male. Thus, men are willing to pay significantly more to participate in an extra round than women. Not being able to control for gender could thus explain why the treatment differences for the first measure are not statistically significant when performing non-parametric tests.³⁰ The coefficient of the

²⁹The two measures at the extensive margin are positively correlated (0.380), as are performance and supply at the intensive margin (0.804). The four ‘cross-correlations’ are lower, varying between 0.125, and 0.244.

³⁰While the randomization generally balances characteristics like gender well, it does not balance them perfectly. In this experiment, the number of male subjects per treatment varies between about one half and about two thirds. More on the sample characteristics can be found in Table 2.13 in Appendix 2.C.1.

interaction term is not significantly different from zero.

Table 2.8: Regression results for labor supply at the extensive margin, measure 1

	Coefficient	Std. error	Coefficient	Std. error
Tax (1=income tax)	116.52*	(62.75)	125.77**	(62.39)
PG (1=public good)	-7.64	(62.74)	12.42	(61.95)
Tax * PG	-48.57	(88.59)	-85.91	(87.84)
Intercept	1398.48***	(44.38)	1378.04***	(220.72)
Age			0.02	(9.37)
Gender (1=male)			142.80***	(45.28)
Studies (1=econ+science)			10.90	(49.92)
Lab experience (1=yes)			-84.08	(67.99)
Observations	240		240	

Notes: The table shows outcomes of a Tobit regression. Individual outcomes are integers between 1056 and 2400. 20 observations are left censored and 4 are right censored. *, **, and *** stand for significance at the 10-, 5-, and 1-percent level.

Table 2.9: Regression results for labor supply at the extensive margin, measure 2

	Coefficient	Std. error	Coefficient	Std. error
Tax (1=income tax)	87.94	(77.74)	94.51	(79.18)
PG (1=public good)	-44.64	(78.06)	-35.93	(78.16)
Tax * PG	27.17	(109.82)	-3.24	(110.95)
Intercept	1218.32***	(55.38)	1282.63***	(281.23)
Age			-1.09	(11.95)
Gender (1=male)			96.51*	(57.70)
Studies (1=econ+science)			-77.71	(64.51)
Lab experience (1=yes)			-48.44	(85.97)
Observations	200		200	

Notes: The table shows outcomes of a Tobit regression. Individual outcomes are integers between 1056 and 2400. 42 observations are left censored and 6 are right censored. *, **, and *** stand for significance at the 10-, 5-, and 1-percent level.

For the second measure (where statistical power is a bit lower, because it was elicited only for employees), the sign of the income tax coefficient is positive, but not significantly different from zero. The coefficient of the interaction term between tax and public good condition is insignificant and close to zero. Note that in this case, an outcome that is statistically significant when conducting a Wilcoxon rank sum test (measure 2 between the public good treatments) loses its statistical significance in a regression setting.

Looking at the overall evidence on labor supply at the extensive margin, there is some evidence that this labor supply is higher under an income tax than under an employer payroll tax. The evidence is weaker than for the findings on public sector size preferences and subjective well-being. However, the fact that the measures we extracted are higher in all four cases under an income tax and the fact that a part of these differences is statistically significant with non-parametric tests while another part is significant in a regression analysis where one can control for observed covariates (in particular gender) gives support to the claim of higher labor supply under an income tax. As for the three mechanisms, gross wage illusion seems to be dominant when participants make their decisions on labor supply at the extensive margin, because the higher gross wage in the income tax case makes subjects more willing to work. The tax loss effect and warm glow seem to play at most a minor role (the tax loss effect weakens the effect of gross wage illusion; warm glow would lead to a positive interaction term between income tax and public good).

Labor Supply at the Intensive Margin and Job Performance

Table 2.10 gives an overview of the outcomes on labor supply at the intensive margin and job performance. Both measures are larger under a payroll tax when the tax proceeds are wasted, but larger under an income tax when tax proceeds are invested in a profitable public good (the differences in the nothing in return treatments are very small, however). Only one of the findings is (marginally) statistically significant.

Table 2.10: Overview and tests, labor supply at the intensive margin and job performance

	Employer payroll tax mean (std. error)	Income tax mean (std. error)	Treatment diff. p-value Wilcoxon
Intensive margin			
Nothing in return	1570.5 (58.1)	1522.5 (66.2)	0.703
Public good	1426.5 (66.7)	1570.5 (60.6)	0.072
Job performance			
Nothing in return	20.56 (1.07)	19.64 (1.17)	0.679
Public good	18.40 (1.21)	21.02 (1.23)	0.139

Notes: Individual outcomes are between 0 and 1920 for labor supply at the intensive margin and positive integers for job performance (the maximum achieved in the data is 37). Larger numbers represent higher labor supply and better job performance. The p-values stem from two-sided Wilcoxon rank sum tests.

Tables 2.11 and 2.12 show the regression outcomes for labor supply at the intensive margin and job performance. For both outcome variables (and independently of whether or not covariates are included), the coefficient of the income tax dummy is negative, but small

and insignificant. For both variables, the coefficient of the interaction of tax and public good condition is positive (it is marginally significant for labor supply at the intensive margin and for both variables much larger than the negative coefficient of the tax type dummy). This hints slightly at larger labor supply at the intensive margin and job performance under an income tax when the tax proceeds are used beneficially. Not surprisingly, subjects with a relatively high math content in their studies (economics and science) are better at solving the problems than other students – the coefficient of the dummy indicating the field of study in Table 2.12 is positive and significantly different from zero.

Table 2.11: Regression results for labor supply at the intensive margin

	Coefficient	Std. error	Coefficient	Std. error
Tax (1=income tax)	-78.59	(108.28)	-86.45	(111.44)
PG (1=public good)	-190.85*	(107.92)	-198.75*	(108.42)
Tax * PG	263.98*	(153.21)	281.49*	(155.95)
Intercept	1658.14***	(77.35)	1636.92***	(390.39)
Age			1.61	(16.81)
Gender (1=male)			-21.96	(80.57)
Studies (1=econ+science)			143.60	(89.72)
Lab experience (1=yes)			-107.49	(121.60)
Observations		199		199

Notes: The table shows outcomes of a Tobit regression. Individual outcomes are between 0 and 1920. No observations are left censored, 41 are right censored. *, **, and *** stand for significance at the 10-, 5-, and 1-percent level.

Table 2.12: Regression results for job performance

	Coefficient	Std. error	Coefficient	Std. error
Tax (1=income tax)	-0.87	(1.67)	-0.53	(1.67)
PG (1=public good)	-2.22	(1.67)	-1.87	(1.64)
Tax * PG	3.55	(2.37)	3.14	(2.34)
Intercept	20.51***	(1.18)	18.91***	(5.88)
Age			-0.04	(0.25)
Gender (1=male)			1.87	(1.21)
Studies (1=econ+science)			3.20**	(1.35)
Lab experience (1=yes)			-1.13	(1.81)
Observations		199		199

Notes: The table shows outcomes of a Tobit regression. Individual outcomes are positive integers. 4 observations are left censored. *, **, and *** stand for significance at the 10-, 5-, and 1-percent level.

Examining the evidence on intensive labor supply and job performance, we are mainly left with a null-result. The mechanisms we discussed seem to be either balanced (gross wage illusion and the tax loss effect work in opposite directions) or not strong enough when subjects take these decisions ‘under stress’ (i.e. while working). The fact that the interaction term of tax and public good condition is positive for both outcome variables and marginally significant for labor supply at the intensive margin is a further (albeit relatively weak) indication that our mechanisms play a role in subjects’ decisions.

2.6 Discussion

The question whether a labor tax levied at the employees’ side of the labor market and one levied at the employers’ side are equivalent is relevant for policy making, political economics and optimal taxation theory. In this chapter we have investigated this LSE in a controlled laboratory experiment. Specifically, we have focused on the effects of this distinct framing of otherwise equivalent taxes, arguing that framing effects provide direct evidence against tax equivalence. Our results support the claim that these duties are not equivalent. In particular, our results suggest that employees prefer a larger public sector and that subjective well-being is higher when the tax is levied on the employer’s side, while labor supply tends to be lower at the extensive margin. We have also highlighted three mechanisms that may explain the framing effects we observe, to wit, gross wage illusion, tax loss effect, and warm glow. Our results show that all three seem to be at work, and have distinct effects on preferred public sector size, subjective well-being and labor supply.

The policy implications of our results are non-trivial as they show that who formally pays the tax affects individuals in different ways. If citizens’ subjective well-being is the government’s main concern, the policy implications of our findings are relatively straightforward. We observe that subjective well-being is higher under an employer payroll tax. This finding thus suggests that, *ceteris paribus*, it would be better to levy taxes on the employer’s side. Concerning labor supply, most economists would probably agree that taxes should be chosen that minimize the disruption of the price mechanism in the labor market. As labor taxes are generally thought to reduce labor supply, this implies that the tax should be chosen that maximizes labor supply. If people work more under one tax scheme than under the other, using this scheme brings society closer to first best and thus increases social welfare. The experimental findings concerning labor supply are that people tend to supply more labor under an income tax than under an employer payroll tax (at the extensive margin; there are indications of similar effects at the intensive margin and performance, see Fochmann et al., 2010b). Thus, the optimal policy for the government when considering only labor supply

responses would be to rely only on income taxes. It is less straightforward to draw policy conclusions from our results regarding public sector size preferences. To start, note that such preferences may be expressed in votes and therefore may affect the actual size of the public sector in the economy. If one assumes that there is something like an optimal size, it seems that the combination of labor taxes is optimal that induces individuals to prefer this size of the public sector. However, while it is already difficult in general to derive the optimal size of the public sector in this experiment, one should not make inferences from this experiment about the distance to the optimum in the real world (see Footnote 26).

The aspects we investigated lead, when considered separately, to distinct implications and it is quite possible that the optimal policy involves a mix of labor taxes levied at the employers' and employees' sides. This could explain the 'puzzle' of why these theoretically equivalent duties often exist side by side instead of governments simply adopting the duty that minimizes collection and compliance costs, even in the absence of labor market frictions. An alternative solution to this 'puzzle' based on political-economic considerations is that politicians choose a tax mix that induces the public to prefer a size of the public sector close to the one the politicians themselves favor.

Our results share some features with those reported by others. Sausgruber and Tyran (2005, 2011) and our research have in common that an individual's willingness to pay taxes is larger if these taxes are not levied on this individual directly (thus when individuals irrationally feel that they do not pay the tax). A further commonality is that the liability side of taxes seems to have a much stronger effect on political preferences and voting decisions than on 'hard' economic decisions (such as labor supply or market behavior in a buyer-seller environment). With Blumkin et al. (2012) and Fochmann et al. (2010b) this research shares the finding that labor supply tends to be higher when taxes are indirect or less salient (although our findings on labor supply are admittedly weaker than our findings on the other dimensions).

A potential drawback of laboratory experiments is a limited external validity due to the artificiality of the setting or the subject pool used. There are, of course, no guarantees that subjects' choices in a laboratory experiment fully reflect their behavior outside of the laboratory; one could easily argue that no experiment, no theory, and no econometric field data analysis fully reflects real-world behavior. In fact, the assumptions needed to generalize laboratory findings to the field are precisely the same as those needed to establish the generalizability of results obtained by using observational field data (Falk and Heckman, 2009). Moreover, laboratory experimentation provides some advantages that are much more difficult to achieve otherwise. For example, knowledge about causal mechanisms is often crucial for understanding the world outside the laboratory. This requires controlled variation and the laboratory allows for such control in a way that is generally impossible outside of the

laboratory.

In addition, with a careful experimental design one can overcome problems of generalizability and gain knowledge about specific features that are relevant for the real world. Problems with external validity can to a large degree be avoided if one focuses on directions of treatment effects rather than on levels and quantifications (see, e.g., Charness and Kuhn, 2011). This is the approach we have taken. As long as general trade-offs are concerned (e.g. a trade-off between earning money and experiencing disutility from working, a trade-off that is commonly assumed in labor economics and that we have induced in the laboratory) and as far as only directions of treatment effects are concerned, the results tend to be robust and can be expected to carry over to the real world (see Schram, 2005, and the references therein). In comparison, claims on quantities and levels should be made cautiously – in fact, we discourage the reader from making such claims based on this research (for example, we refrain from making claims such as that a y -percent increase of some variable could be achieved by a shift in the liability side of the tax). Finally, it goes without saying that our understanding of liability-side (non-)equivalence would benefit from laboratory evidence being confirmed or challenged by further studies involving other methodologies.

Summing up, it seems clear from our results that boundedly rational behavior plays an important role in individuals' reactions to taxes. Classic optimal taxation theory, which assumes individual rationality, is thus based on an empirically shaky foundation. The development of normative optimal taxation models encompassing non-rational perception of taxes seems an important line of future research. The same holds for the incorporation of tax perception into positive models of political economics. Our results may aid in the development of such models.

Appendix 2.A Instructions, Test Questions, and Questionnaire

We reproduce here the on-screen instructions and test questions. They are in the same order as they appear during the experiment (in between are the tasks as described in Section 2.3). We have split the instructions for the different treatments where relevant. Aside from the on-screen instructions, subjects received a one-page summary which has no new information. Text that is bold and/or italic on the screen is also so in this appendix. Multiple choice test questions are shown here with empty squares for the possible answers, test questions where an input is required are shown with an empty circle below the question. Note that it is only possible to proceed from a page of test questions after all questions on this page have been answered correctly. If not, the message “You did not answer all questions correctly. Take another look at the instructions or raise your hand if you need help.” is displayed, without telling the participant which question(s) has/have been answered incorrectly. It is thus necessary for the participants to fully understand the instructions and to know the answers to the questions rather than just clicking through them, also for multiple choice questions. In Section 2.A.2 we reproduce the questionnaire for the subjects in the role of employees.

2.A.1 Instructions and Test Questions

FIRST INSTRUCTION SCREEN, ALL TREATMENTS

Welcome to this experiment!

Depending on your decisions and the decisions of other participants in today’s experiment, you can earn money. You will be paid privately at the end of the experiment. Your earnings will not be revealed to anyone. This is an anonymous experiment; your decisions will only be linked to your station id and not to your name in any way. The experiment will take approximately 2 hours.

This experiment may involve gains and losses. Whether the possibility of a loss occurs is completely determined by your own decisions. It is thus possible (though unlikely) that you make a negative amount of money in the experiment. In this case, your losses will be deducted from your earnings and from the show-up fee.

This experiment is composed of different parts. You will receive instructions for the different parts before they begin. After the instructions you will have to answer a few test questions before you can continue.

Please read all instructions carefully. You are not allowed to speak with other participants or to communicate with them in any other way.

Payments in most parts are in points, but there are also payments in Euro. At the end of the experiment points will be exchanged into Euro at the exchange rate of 600 points = 1 Euro.

By showing up, you have already earned 7 Euro. This show-up fee will be added to your earnings from the experiment.

If you want to ask a question, please raise your hand and someone will come to your desk.

SECOND INSTRUCTION SCREEN, ALL TREATMENTS

In this part you will be randomly divided into groups of six people each. One person in each group will be randomly determined to be "employer" the other five persons will be "employees". The group composition will not change during the whole experiment.

This part consists of four rounds. Each will last for 8 minutes. In each round, the employer hires the five employees to perform a work task.

What the employees have to do:

You (employees only) will see two matrices on the screen. Each matrix has 10 rows and 10 columns and is filled with randomly generated numbers. Your job is to find the largest number in each of the matrices and then to add them up. You are allowed to use the pocket calculators on your table. For each correct solution, the employer will pay you a wage. This wage becomes lower for each new problem you face.

After entering your answer you will be told whether your answer is correct or not (please note that the time will continue to run while you see this result). Subsequently, irrespective of whether your answer is correct or incorrect, a new pair of matrices will appear. This means that for each pair, you have only one attempt to provide the correct answer.

Instead of trying to solve problems you can also choose to use a "fixed payment option". You will see a button saying "Go to fixed payment option" at the bottom right of your screen. If you click on this button you will receive a fixed payment for each second remaining in the round; while the time ticks away you will see a basically empty screen. You will then not be able to solve any more problems in this round (thus you cannot go back and forth between problem solving and the fixed payment option).

At the top of the screen you can see how many of your answers in this round were right and wrong. Here, you can also see your total wage for this round. You will see the time that remains in this round in the upper right corner of the screen. You will also always see

the wage you will receive if your next problem is solved correctly. You can try to solve at most thirty of these problems per round (which will be much more than anyone can actually solve).

What the employers have to do: While the employees are working, the employers do not need to do anything.

After each round there will be a short one-click questionnaire. Please answer this question as honestly and accurately as you can.

FIRST PART OF THE THIRD INSTRUCTION SCREEN

TREATMENT *EN*

Payments:

For each correct addition an employee makes, his/her employer (the employer in the same group) will receive a certain number of points. From these points, the employer pays a wage to the employee. Aside from the wage paid to the employee, the employer will also pay (to the experimenter) an amount equal to 66.7% of the employee's wage. The employer thus has to pay more than only the employees' wages.

The wage an employee receives for a correct addition depends on how many problems this employee has already attempted to solve in this round. If the first attempt is correct, the employee's wage is 280.8 points. For each subsequent attempt the wage decreases by 23.4 points, but it is never smaller than zero. As employee you will always see on the screen the wage you will receive if the next addition is correct. For an incorrect addition neither the employer nor the employee will receive any points. The amount the employer receives for a correct addition of an employee is such that he/she will in the end always keep as a profit 49.8 points per correct addition.

Instead of solving problems, an employee can also go to the fixed payment option. If an employee goes to the fixed payment option, he/she receives 2.2 points per second that is still remaining in this round. The employer earns nothing from the fixed payment option, he/she only earns points for correctly solved problems.

TREATMENT *IN*

Payments:

For each correct addition an employee makes, his/her employer (the employer in the same group) will receive a certain number of points. From these points, the employer pays a wage to the employee. From the employee's wage, 40% will be deducted, such that an employee

in the end only receives 60% of his/her wage.

The wage an employee receives for a correct addition depends on how many problems this employee has already attempted to solve in this round. If the first attempt is correct, the employee's wage is 468 points. For each subsequent attempt the wage decreases by 39 points, but it is never smaller than zero. As employee you will always see on the screen the wage you will receive if the next addition is correct. For an incorrect addition neither the employer nor the employee will receive any points. The amount the employer receives for a correct addition of an employee is such that he/she will in the end always keep as a profit 49.8 points per correct addition.

Instead of solving problems, an employee can also go to the fixed payment option. If an employee goes to the fixed payment option, he/she receives 2.2 points per second that is still remaining in this round. From these points nothing is deducted. The employer earns nothing from the fixed payment option, he/she only earns points for correctly solved problems.

TREATMENT *EP*

Payments:

For each correct addition an employee makes, his/her employer (the employer in the same group) will receive a certain number of points. From these points, the employer pays a wage to the employee. Aside from the wage paid to the employee, the employer will also pay an amount that equals 66.7% of the employee's wage. The proceeds will be put into a common fund. Each group has its own common fund. At the end of the experiment, the total number of points in the common fund will be multiplied by 1.3 and then distributed equally among all employees in this group.

The wage an employee receives for a correct addition depends on how many problems this employee has already attempted to solve in this round. If the first attempt is correct, the employee's wage is 239.32 points. For each subsequent attempt the wage decreases by 19.94 points, but it is never smaller than zero. As employee you will always see on the screen the wage you will receive if the next addition is correct. For an incorrect addition neither the employer nor the employee will receive any points. The amount the employer receives for a correct addition of an employee is such that he/she will in the end always keep as a profit 49.8 points per correct addition.

Instead of solving problems, an employee can also go to the fixed payment option. If an employee goes to the fixed payment option, he/she receives 2.2 points per second that is still remaining in this round. The employer earns nothing from the fixed payment option, he/she

only earns points for correctly solved problems (thus, the employer will not contribute points to the common fund for the time an employee uses the fixed payment option).

TREATMENT *IP*

Payments:

For each correct addition an employee makes, his/her employer (the employer in the same group) will receive a certain number of points. From these points, the employer pays a wage to the employee. From the employee's wage, 40% will be deducted and put into a common fund. Each group has its own common fund. At the end of the experiment, the total number of points in the common fund will be multiplied by 1.3 and then distributed equally among all employees in this group.

The wage an employee receives for a correct addition depends on how many problems this employee has already attempted to solve in this round. If the first attempt is correct, the employee's wage is 398.87 points. For each subsequent attempt the wage decreases by 33.24 points, but it is never smaller than zero. As employee you will always see on the screen the wage you will receive if the next addition is correct. For an incorrect addition neither the employer nor the employee will receive any points. The amount the employer receives for a correct addition of an employee is such that he/she will in the end always keep as a profit 49.8 points per correct addition.

Instead of solving problems, an employee can also go to the fixed payment option. If an employee goes to the fixed payment option, he/she receives 2.2 points per second that is still remaining in this round. From these points nothing is deducted. The employer earns nothing from the fixed payment option, he/she only earns points for correctly solved problems.

SECOND PART OF THE THIRD INSTRUCTION SCREEN, ALL TREATMENTS

Before the rounds start:

Before the problem solving rounds start you will be told whether you are employer or employee. Before this, you will be asked how much you would be willing to pay in order to work (and get paid) for an extra round in case that you are employee. There will be a random mechanism deciding whether or not this extra round will actually take place. If the round does not take place or if you are employer, the number you state will not have any consequences for you. This extra round will take place (if it does take place) at the end of the experiment. It will be the same as the regular rounds, just the number of people that work in your group may be different, simply because some may not be working in this extra round. Also the payments will be the same as in the regular rounds.

If the random process we use determines that the extra round will in fact take place, the following occurs. The price of participating in this extra round is randomly determined and lies between 1056 and 2400 points. If this price is lower than the number you state, the price will be deducted from your earnings and you will work and get paid for an extra round. If the price is higher than the number you state, nothing will be deducted from your earnings and you will not work for an extra round. Note that we will not pay anyone until the whole experiment is finished. Note that you will have to state an amount of at least 1056 points (this is the amount of points you get for a round if you immediately switch to the fixed payment option).

Thus, you will not pay for the extra round the amount you state that you are willing to pay. You (if you are employee) either pay the randomly determined price – in case that this price is lower than the number you state – or you pay nothing and do not play an extra round.

If you are employer you will again not have to do anything in the extra round, but you will earn the same amount of points per correct addition of your employees as in the regular rounds.

FIRST TEST QUESTION SCREEN, ALL TREATMENTS

Before the experiment starts, we will ask you some questions to check your understanding. You can return to the instructions by clicking on the menu at the top of the screen.

When do the matrices that you see on the screen change?

- After you have entered the correct solution.
- After you have entered a number, irrespective of whether it is correct or not.
- After you have entered the correct solution or after 1 minute.

How often can you go back from the fixed payment option to solving problems in one round?

- Never.
- Once.
- As often as you like.

How many minutes does each of the rounds last?

○

The matrices that you will see during the experiment will be much larger. For now, assume that one of the matrices on your screen consists of the numbers 19, 23, 41, 16, 25, 30, 12,

29, 22 and the other matrix consists of the numbers 31, 36, 20, 15, 28, 38, 17, 19, 31. What would be the correct number to enter?

○

How large will each of the matrices that you see during the experiment be?

- 5 rows and 5 columns (25 numbers)
- 10 rows and 10 columns (100 numbers)
- 20 rows and 20 columns (400 numbers)

FIRST PART OF THE SECOND TEST QUESTION SCREEN

TREATMENT *EN*

Imagine that you are employee. On top of your wage for each correct addition, how much extra will be taken away from your employer (expressed as a percentage of your wage)?

- 33.3
- 66.7
- 50

TREATMENT *IN*

Imagine that you are employee. How much of your wage for each correct addition will be taken away from you (as a percentage of your wage)?

- 22
- 40
- 50

TREATMENT *EP*

Imagine that you are employee. On top of your wage for each correct addition, how much extra will be taken away from your employer and contributed to the common fund (expressed as a percentage of your wage)?

- 33.3
- 66.7

- 50

TREATMENT *IP*

Imagine that you are employee. How much of your wage for each correct addition will be taken away from you and contributed to the common fund (as a percentage of your wage)?

- 22
- 40
- 50

SECOND PART OF THE SECOND TEST QUESTION SCREEN, ALL TREATMENTS

If you are employer, how many points will you earn for each second that one of your employees chooses the fixed payment option?

-

If you are employee, how many points will you earn per *minute* that you choose the fixed payment option?

-

THIRD TEST QUESTION SCREEN, ALL TREATMENTS

As you know, before you will start the experiment you will be asked how much you would be willing to pay to participate in an extra round. The test questions now concern this statement. Please note that any numbers here are only meant to serve as an example. The content is not informative on what you should decide in the experiment.

Imagine that you state an amount of 1300 points and that the randomly drawn price is 1344 points. What will happen?

- 1300 points are deducted from your earnings, but you will not participate in the extra round.
- 1344 points are deducted from your earnings and you will participate in the extra round, if it takes place.
- 1300 points are deducted from your earnings and you will participate in the extra round, if it takes place.

- No points are deducted from your earnings and you will not participate in the extra round, if it takes place.

Imagine that the randomly drawn price is 1257.3 points and that it is lower than the amount you state. Imagine that the extra round takes place and the you go immediately to the fixed payment option. Will you then earn more in the extra round than the price you pay?

- Yes.
 No.

Imagine that the amount you state is 1100 points and the randomly determined price is 1429 points. Will you be able to leave the experiment earlier in this case than if you had stated 2000 points?

- Yes.
 No.
 That depends on the choices of the other participants in my group.

SUMMARY OF THE INSTRUCTIONS FOR THE FIRST PART, ALL TREATMENTS

Summary of instructions

Please note that you have a short summary of instructions including the payments lying on your desk.

You will be randomized into employees and employers.

If you are employee you solve problems that consist of finding the largest number in each of two matrices and adding the numbers up. You will receive a wage for the correct additions. Instead of working you can also choose a fixed payment option. For the payments see the sheet on your desk.

If you are employer you hardly have to do anything.

Before beginning you will be asked how much you would at most be willing to pay to participate in an extra round at the end of the experiment if you are employee. You will not pay the amount you enter, but either a randomly drawn price if this price is lower than the amount you enter or nothing.

If you want to ask a question, please raise your hand and someone will come to your desk.

INSTRUCTIONS AFTER THE REGULAR WORK ROUNDS

TREATMENTS *EN* AND *IN*

Welcome to the next part of the experiment

There is again a chance to participate in an extra round at the end of the experiment. The round will again be the same as the regular work rounds. Also the payments will be the same.

You will be asked to state the maximum price you are willing to pay to participate in this extra round. If the randomly drawn price (between 1056 and 2400) is lower than what you stated you pay this price and participate in the extra round. (The rules are thus the same as in the beginning of the experiment.)

Please note that you will not be able to finish the experiment early if you do not participate in the extra round. Please also note that it is uncertain whether or not the extra round will take place. In fact, either the extra round based on your maximum price decision before the regular rounds will take place or an extra round based on your next maximum price decision will take place. Therefore, if an extra round will take place based on your statement from before the regular rounds, this extra round based on the statement now will not take place and if no extra round will take place based on your statement from before the regular rounds, this extra round based on the statement now will take place.

This means that from your two maximum prices, only one will be considered.

If you have any questions, please raise your hand and one of us will come to your desk to answer the question.

TREATMENTS *EP* AND *IP*

Welcome to the next part of the experiment

There is again a chance to participate in an extra round at the end of the experiment. The round will again be the same as the regular work rounds. Also the payments will be the same (the common fund will be distributed among all employees of the group, also among the ones that do not participate).

You will be asked to state the maximum price you are willing to pay to participate in this extra round. If the randomly drawn price (between 1056 and 2400) is lower than what you stated you pay this price and participate in the extra round. (The rules are thus the same as in the beginning of the experiment.)

Please note that you will not be able to finish the experiment early if you do not participate

in the extra round. Please also note that it is uncertain whether or not the extra round will take place. In fact, either the extra round based on your maximum price decision before the regular rounds will take place or an extra round based on your next maximum price decision will take place. Therefore, if an extra round will take place based on your statement from before the regular rounds, this extra round based on the statement now will not take place and if no extra round will take place based on your statement from before the regular rounds, this extra round based on the statement now will take place.

This means that from your two maximum prices, only one will be considered.

If you have any questions, please raise your hand and one of us will come to your desk to answer the question.

INSTRUCTIONS BEFORE THE PREFERENCE FOR PUBLIC SECTOR SIZE ELICITATION

TREATMENT *EN*

Welcome to the next part of the experiment

There will be one final extra round in the end of the experiment. Everyone will participate in this final round. This round will be a bit different from the regular rounds.

In this round, the amount that the employer pays beside the wage is no longer paid to the experimenter. Instead, it is put into a common fund. Each group has its own common fund. At the end of the experiment, the total number of points in the common fund will be multiplied by a certain factor and then distributed equally among all employees in this group.

Some of the parameters will now be decided on by the employees. You will be asked to state which set of parameters you would prefer for this extra round of the experiment. Later, the preferred set of parameters of one employee of your group will be chosen randomly and the additional extra round will take place with this set of parameters.

You will see a slider (a button that you can move horizontally) to determine your preferred set of parameters. If you move the slider from left to right, the factor with which the points in the common fund will be multiplied decreases. At the same time, the number of points that your employer contributes to the common fund increases. You are asked to choose the position of the slider that you prefer. Please note that the profit of the employer is held constant. Therefore, higher contributions to the common fund will lead to lower wages.

You can look at this problem in the following way (although this is not the only possible way). You would like the multiplication factor to be big, because it will be multiplied with

the points in the common fund before these points are distributed among the employees of your group. But if you choose the slider to be as far to the left as possible, nothing will be contributed to the common fund at all and no one can profit from the large multiplication factor. As explained, you should choose the combination of the multiplication factor and the size of the contributions to the common fund that you prefer most.

If you have any questions, please raise your hand and one of us will come to your desk.

TREATMENT *IN*

Welcome to the next part of the experiment

There will be one final extra round in the end of the experiment. Everyone will participate in this final round. This round will be a bit different from the regular rounds.

Now the points that are deducted from your wage are no longer "lost". Instead, they are put into a common fund. Each group has its own common fund. At the end of the experiment, the total number of points in the common fund will be multiplied by a certain factor and then distributed equally among all employees in this group.

Some of the parameters will now be decided on by the employees. You will be asked to state which set of parameters you would prefer for this extra round of the experiment. Later, the preferred set of parameters of one employee of your group will be chosen randomly and the additional extra round will take place with this set of parameters.

You will see a slider (a button that you can move horizontally) to determine your preferred set of parameters. If you move the slider from left to right, the factor with which the points in the common fund will be multiplied decreases. At the same time, the number of points that is deducted from your wage and contributed to the common fund increases. You are asked to choose the position of the slider that you prefer.

You can look at this problem in the following way (although this is not the only possible way). You would like the multiplication factor to be big, because it will be multiplied with the points in the common fund before these points are distributed among the employees of your group. But if you choose the slider to be as far to the left as possible, nothing will be contributed to the common fund at all and no one can profit from the large multiplication factor. As explained, you should choose the combination of the multiplication factor and the size of the contributions to the common fund that you prefer most.

If you have any questions, please raise your hand and one of us will come to your desk.

TREATMENT *EP*

Welcome to the next part of the experiment

There will be one final extra round in the end of the experiment. Everyone will participate in this final round. This round will be a bit different from the regular rounds.

Some of the parameters will now be decided on by the employees. You will be asked to state which set of parameters you would prefer for this extra round of the experiment. Later, the preferred set of parameters of one employee of your group will be chosen randomly and the additional extra round will take place with this set of parameters.

You will see a slider (a button that you can move horizontally) to determine your preferred set of parameters. If you move the slider from left to right, the factor with which the points in the common fund will be multiplied decreases (of course, this refers only to the points contributed to the common fund in this extra round). At the same time, the number of points that your employer contributes to the common fund increases. You are asked to choose the position of the slider that you prefer. Please note that the profit of the employer is held constant. Therefore, higher contributions to the common fund will lead to lower wages.

You can look at this problem in the following way (although this is not the only possible way). You would like the multiplication factor to be big, because it will be multiplied with the points in the common fund before these points are distributed among the employees of your group. But if you choose the slider to be as far to the left as possible, nothing will be contributed to the common fund at all and no one can profit from the large multiplication factor. As explained, you should choose the combination of the multiplication factor and the size of the contributions to the common fund that you prefer most.

If you have any questions, please raise your hand and one of us will come to your desk.

TREATMENT *IP*

Welcome to the next part of the experiment

There will be one final extra round in the end of the experiment. Everyone will participate in this final round. This round will be a bit different from the regular rounds.

Some of the parameters will now be decided on by the employees. You will be asked to state which set of parameters you would prefer for this extra round of the experiment. Later, the preferred set of parameters of one employee of your group will be chosen randomly and the additional extra round will take place with this set of parameters.

You will see a slider (a button that you can move horizontally) to determine your preferred

set of parameters. If you move the slider from left to right, the factor with which the points in the common fund will be multiplied decreases (of course, this refers only to the points contributed to the common fund in this extra round). At the same time, the number of points that is deducted from your wage and contributed to the common fund increases. You are asked to choose the position of the slider that you prefer.

You can look at this problem in the following way (although this is not the only possible way). You would like the multiplication factor to be big, because it will be multiplied with the points in the common fund before these points are distributed among the employees of your group. But if you choose the slider to be as far to the left as possible, nothing will be contributed to the common fund at all and no one can profit from the large multiplication factor. As explained, you should choose the combination of the multiplication factor and the size of the contributions to the common fund that you prefer most.

If you have any questions, please raise your hand and one of us will come to your desk.

FIRST PART OF THE TEST QUESTION SCREEN BEFORE THE PREFERENCE FOR PUBLIC SECTOR SIZE ELICITATION, ALL TREATMENTS

If you go to the fixed payment option instead of solving problems, will this increase the number of points in the common fund in any way?

- Yes.
- No.
- That depends on the set of parameters chosen.

If your preferred position of the slider is all the way to the left and your choice is the one that is randomly picked, how many points will you receive for each correct addition by one of the other employees in your group?

o

SECOND PART OF THE TEST QUESTION SCREEN BEFORE THE PREFERENCE FOR PUBLIC SECTOR SIZE ELICITATION

TREATMENTS *EN* AND *EP*

Does the gross number of points that your employer receives for each correct addition (i.e. the number before he/she pays the wages and the contributions to the common fund) depend on the set of parameters chosen?

- Yes.
- No.

TREATMENTS *IN AND IP*

Does the gross number of points that your employer receives for each correct addition (i.e. the number before he/she pays the wages) depend on the set of parameters chosen?

- Yes.
- No.

THIRD PART OF THE TEST QUESTION SCREEN BEFORE THE PREFERENCE FOR PUBLIC SECTOR SIZE ELICITATION, ALL TREATMENTS

Does the profit your employer keeps in the end from each correct addition by one of his/her employees depend on the set of parameters chosen?

- Yes.
- No.

INTRODUCTION OF THE LOTTERIES CONCERNING THE LOSS AVERSION ELICITATION, ALL TREATMENTS

Welcome to the next part of the experiment

In this part you will be presented lotteries A and B.

*A: Play the following lottery **once**.*

Win 4 Euro with probability one half, lose 2.5 Euro with probability one half.

*B: Play the following lottery **six times in a row**.*

Win 4 Euro with probability one half, lose 2.5 Euro with probability one half.

You can state for each of the two lotteries whether or not you would like to participate in them. Only one of the two lotteries will actually be played out, and it will only be played out for you if you decided to participate in it. It will later be randomly decided whether lottery A or lottery B will be played out.

If you decided to participate in the lottery that is chosen, it will be played out and you will receive the earnings. If you decided not to participate in this lottery you will neither earn nor lose any money.

If you have any questions, please raise your hand and one of us will come to your desk.

TEST QUESTIONS BEFORE CHOOSING THE LOTTERIES, ALL TREATMENTS

Please answer the following questions. (Please note that you should state decimals with a point, not with a comma.)

If you choose not to participate in either of the lotteries, how much money will you earn?

- Euro

If you state that you would like to participate in both lotteries and lottery A is chosen randomly to be played out, how much money can you lose in the worst case scenario?

- Euro

If you state that you would like to participate in lottery B and lottery B is chosen randomly to be played out how much money can you win in the best case scenario?

- Euro

2.A.2 Questionnaire

ALL TREATMENTS

Questionnaire

The experiment has almost ended. Please fill out the following questionnaire. When you have finished, you will return to the summary of your earnings. When everyone has finished with the questionnaire we will start paying you and you may leave.

Gender:

- Male
- Female

Age:

-

Have you participated in a CREED experiment before?

- No
- Yes, once
- Yes, more than once

Department where you study:

- UVA – Faculty of Economics and Business
- UVA – Faculty of Social and Behavioural Sciences - Psychology
- UVA – Faculty of Social and Behavioural Sciences - non psychology
- UVA – Faculty of Science
- UVA – IIS: beta gamma bachelor
- UVA – Faculty of Law
- UVA - Faculty of Humanities
- UVA - Faculty of Medicine
- UVA - Faculty of Dentistry
- Another university
- A Dutch 'hogeschool' (HBO)
- Different

Did you sometimes switch to the fixed payment option? If yes, how did you decide when to switch?

○

Were there things you did not understand completely/correctly during the experiment? If yes, please state which parts. You can also leave any other comment here, if there is something you think we might be interested in knowing.

○

Consider the socio-economic system in the Netherlands. What do you think about the public sector? It is...

- Much too big
- Too big

- About right
- Too small
- Much too small
- Don't know

Appendix 2.B Screenshots

Figures 2.5 to 2.7 (depicted at the end of this appendix) show screenshots from the experiment. Figure 2.5 shows a screenshot of an employee's computer during the work task (taken in treatment *IN*). Figure 2.6 shows the subjective well-being self-assessment using the SAM-V-9 self-assessment manikin, and Figure 2.7 shows the slider that is used for the elicitation of preferences concerning the size of the public sector (taken in treatment *EN*).

Amount of correct additions in this round so far: 1

Amount of incorrect additions in this round so far: 1

Remaining problems that can be solved in this round: 28

Round 1 of 4

Total wage earned in this round: 429 (minus 171.6)

Time: 06:53

If your next answer is correct you will receive 390 points (from which 156 will be deducted).

38	39	24	30	58	67	53	60	53	58
73	20	12	79	56	32	29	24	79	40
24	30	43	25	21	59	25	73	78	41
67	23	10	77	35	30	15	25	40	76
48	58	20	36	28	35	15	29	49	57
33	71	30	53	44	72	65	55	56	49
73	62	39	50	78	35	72	22	69	44
54	14	71	50	63	42	71	52	70	17
25	12	13	23	62	68	71	41	65	43
42	16	24	74	38	68	32	56	65	74

83	57	25	13	51	34	66	42	74	63
12	69	29	28	19	37	53	77	57	74
22	22	58	46	75	82	66	82	13	70
14	50	47	58	21	76	56	81	19	30
27	62	52	38	59	36	54	42	68	74
67	21	78	64	25	67	78	77	46	83
20	73	63	13	40	51	45	57	15	45
74	83	81	52	46	71	43	34	42	37
37	62	31	34	13	23	20	78	36	62
54	75	55	67	65	40	59	34	23	28

Largest number in the left matrix plus largest number in the right matrix:

Figure 2.5: Screenshot during a work round

How would you describe your mood at the moment? Please choose according to the shown graph.

The screenshot displays a self-assessment interface. At the top, a question asks the user to describe their mood. Below the question are nine mood icons, each in a square frame and numbered 1 through 9. The icons are arranged in a horizontal row. Below the icons is a vertical list of radio buttons, each corresponding to a number from 1 to 9. To the right of the radio buttons is an 'OK' button.

1
 1

2
 2

3
 3

4
 4

5
 5

6
 6

7
 7

8
 8

9
 9

OK

Figure 2.6: Screenshot during the self-assessment with the SAM-V-9

Instructions Questions **Decision**

Please choose your most preferred position of the slider.

Multiplication factor of the common fund: 3.00
Contribution to the common fund, expressed as a percentage of your wage: 0.00

OK

Figure 2.7: Screenshot of the slider; public sector size preference

Appendix 2.C Sample Composition and Gender Effects

2.C.1 Sample Composition

Table 2.13 shows the composition of the sample in the different treatments of the experiment.

Table 2.13: Sample composition

	<i>EN</i>	<i>IN</i>	<i>EP</i>	<i>IP</i>
Average age	21.6	22.8	21.7	22.0
Gender male	32	25	24	31
Studies econ + science	36	34	34	30
Lab experience	48	44	44	40

Notes: The table shows the composition of the sample. Each of the controls in the regressions in Section 2.5 is considered. The numbers reported are for employees only (the total number of employees in each treatment is 50). The age shown is the average age in the treatment, gender is the number of men, studies is the number of subjects studying economics or science, lab experience shows the number of subjects that had participated in economic experiments before.

2.C.2 An Explorative Analysis of Gender Effects

In the world in- and outside the laboratory, gender differences exist regarding many economic variables (Croson and Gneezy, 2009; for labor supply see e.g. Evers et al., 2008). We therefore explore whether such gender differences exist in our data with respect to the way in which tax framing affects the variables we consider. Note that we are not interested in differences in levels here (e.g., ‘do women produce more output?’), but in differences in the treatment effects (e.g., ‘is female job performance affected differently by tax framing than male job performance?’).

We thus explore whether the results from Section 2.5 are driven by either gender. For this (explorative) analysis of our experimental data, regressions are not optimal, because they would flow over with interaction terms (not only the interaction between gender and tax type would have to be included, but also the interactions between gender and the interaction between tax type and use of tax proceeds as well as the interaction between gender and use of tax proceeds, on top of the ‘regular’ interaction between tax type and use of tax proceeds). Therefore, we only use non-parametric tests here, separately for nothing in return and public good treatments. We split the data according to gender and compare the treatment differences among men and women.

Table 2.14 shows the means of the different outcome variables, split up according to treatment and gender. It is surprising that some of the effects reported above are driven

solely by male participants while some others are driven solely by female participants. Of course, the sample sizes are now much smaller. There are 32 (18) men (women) in treatment *EN*, 25 (25) in *IN*, 24 (26) in *EP*, and 31 (19) in *IP* (cf. Table 2.13).³¹

Table 2.14: Results split up according to gender

	Employer payroll tax mean m/f	Income tax mean m/f	Treatment differences p-value Wilcoxon m/f
<i>ps</i>			
Nothing in return	54.44/45.39	31.84/34.16	0.009/0.146
Public good	40.46/41.85	36.68/27.32	0.634/0.024
<i>sw</i>			
Nothing in return	21.66/22.17	20.71/17.24	0.727/0.004
Public good	22.42/20.00	21.67/22.05	0.814/0.351
<i>le1</i>			
Nothing in return	1459.1/1342.5	1588.5/1448.3	0.105/0.510
Public good	1463.1/1361.3	1507.2/1391.0	0.532/0.751
<i>le2</i>			
Nothing in return	1246.0/1323.4	1406.6/1287.4	0.007/0.726
Public good	1230.3/1228.3	1373.9/1253.1	0.020/0.788
<i>li</i>			
Nothing in return	1534.5/1634.4	1626.7/1418.3	0.353/0.115
Public good	1336.4/1509.8	1627.7/1480.1	0.088/0.333
<i>jp</i>			
Nothing in return	21.06/19.67	22.48/16.80	0.473/0.312
Public good	18.04/18.73	22.53/18.63	0.099/0.881

Notes: The table shows the means of the different outcome variables split according to gender. The p-values stem from two-sided Wilcoxon rank sum tests, comparing treatment effects separately for men and women. We denote by *ps* the preference concerning the size of the public sector; by *sw* the subjective well-being; by *le1* and *le2*, respectively, the first and second measures of labor supply at the extensive margin; by *li* the labor supply at the intensive margin; and by *jp* the job performance.

A first thing to observe is that the aggregate results concerning public sector size preferences are driven by men when there is no public good while they are driven by women when there is a public good. After having previously experienced a regime where tax proceeds are lost, men prefer a significantly larger public sector if the tax (that is then used to

³¹The two observations that have missing data due to computer problems during the experiment are both from male participants (see Footnote 24). The composition of the sample with a bit more men than women has the consequence that there is slightly higher statistical power to detect treatment differences for men than there is for women.

produce a public good) is being levied on the employer. The similar effect for women is smaller. In contrast, when previous experience was with a regime where taxes were already used to produce a public good, men barely respond to differences in who is paying, while women want a much larger public sector if the employer pays the taxes. In terms of subjective well-being, the aggregate result that when proceeds are wasted, people feel better if the tax was paid by the employer, is driven by women. The well-being reported by men barely responds to the framing of the taxes. The labor supply and job performance effects are all mainly driven by men, with highly significant differences (p-values of 0.007 and 0.02) for the second measure of labor supply at the extensive margin (with and without public good, respectively) and marginally significant differences for labor supply at the intensive margin and job performance in the treatments with public good. When it comes down to such labor market responses, women are barely affected by the framing of the tax.

While the results on public sector size preferences are somewhat inconclusive concerning gender effects, the results on subjective well-being and labor supply suggest that gross wage illusion is stronger for men while the tax loss effect is stronger for women. Most of the economic literature looking at gender effects seems to suggest that women are more sensitive to framing than men (see Croson and Gneezy, 2009). In contrast, our results suggest that the mechanisms underlying gender effects in framing differ between men and women. As a consequence, women are more sensitive to framing for some measures while men are more sensitive to framing for others. In this line of reasoning, women are more sensitive concerning subjective well-being where the tax loss effect is the dominant mechanism. On the other hand, men are more sensitive to framing concerning the variables traditionally of most interest to economists – labor supply and job performance where gross wage illusion is more important.

Chapter 3

Monetary Policy under Behavioral Expectations: Theory and Experiment*

3.1 Introduction

Expectations play a crucial role in modern macroeconomic models that are used for scientific research and policy analysis. Usually, these expectations are modeled by assuming a representative fully rational agent. However, the assumption that all agents in an economy are fully rational and able to determine the model consistent expectation of the underlying process governing real-world economic outcomes is highly problematic. A lot of research has shown that humans are generally not able to react fully rationally to the world around them; this research ranges from providing evidence for simple but partly very persistent biases to showing the inability of humans to work with probabilities and to forecast future economic behavior (Tversky and Kahneman, 1974, Grether and Plott, 1979, Tversky and Thaler, 1990, and Kahneman et al., 1991, among many others; see Camerer et al., 2011 for a more recent overview). Also the claim based on evolutionary arguments that behavior deviating from the homogeneous rational expectations solution will be driven out of markets over time has been shown not to be generally true (Brock and Hommes, 1997, 1998, De Grauwe, 2012a; see also Arthur et al., 1997).

We replace the assumption of rational expectations in a macroeconomic model by the assumption that expectations are formed according to a behavioral heuristic switching model, which we take from earlier work. This particular behavioral model of expectation formation has been developed over a long period of time in which (mainly microeconomic) research has been conducted to investigate the questions of how people form expectations and of how they adapt their ways of forming expectations when confronted with observed economic

*This chapter is based on joint work with Cars Hommes and Domenico Massaro (Hommes et al., 2015).

outcomes (see Brock and Hommes, 1997, 1998, Carroll, 2003, Mankiw et al., 2003, Frankel and Froot, 1987, 1991, Bloomfield and Hales, 2002, Branch, 2004, Hommes, 2011, and Assenza et al., 2014b).

A key difference in outcomes between the macroeconomic models with rational and behavioral expectations concerns price stability, i.e. inflation volatility. Assuming rational expectations, there is a clear trade-off for a central bank between fighting inflation volatility and output gap volatility. If the central bank reacts to the output gap in addition to inflation this will under rational expectations always result in an increase of inflation volatility. This is different under behavioral expectations. Starting from a situation in which the central bank does not react to the output gap at all, the central bank can simultaneously decrease inflation volatility and output gap volatility by reacting to the output gap. However, inflation volatility as a function of the extent of output gap reaction is U-shaped. This means that reacting to the output gap on top of inflation will only lower inflation volatility up to a certain level after which inflation volatility starts to increase again.

These different outcomes regarding inflation volatility can be tested in the laboratory. We design a learning-to-forecast experiment where the only difference between treatments consists in the monetary policy rule used by the central bank. In one treatment the central bank only reacts to inflation, in the other it additionally reacts to the output gap. Our experimental results support the claim that inflation volatility can be lowered when the central bank also reacts to the output gap, in line with the predictions of the behavioral model.³²

Our results from the behavioral model and the experimental data have clear policy implications for central banks with the sole aim of achieving price stability such as the European Central Bank.³³ Even if these banks only care about price stability, this goal is better achieved if they also react to changes in the output gap. This is important and at odds with standard macroeconomic thinking built upon full rationality.

This chapter is organized as follows. In Section 3.2 we describe how we model the economy and the formation of expectations. We also show the main differences between the

³²This research builds upon various streams of literature; in particular on the literature on experimental macroeconomics and learning-to-forecast experiments (e.g. Marimon and Sunder, 1993, Van Huyck et al., 1994, Bernasconi and Kirchkamp, 2000, Kelley and Friedman, 2002, Lei and Noussair, 2002, Arifovic and Sargent, 2003, Hommes et al., 2005b, Adam, 2007, Heemeijer et al., 2009, Davis and Korenok, 2011, Bao et al., 2012, Kryvtsov and Petersen, 2013, Cornand and M'Baye, 2013, Pfajfar and Zakelj, 2014, Assenza et al., 2014b; see Duffy, 2012, and Assenza et al., 2014a, for surveys) and on the literature on behavioral macroeconomics (in particular works that consider monetary and fiscal policy when allowing for a departure from the hypothesis of rational expectations; e.g. Bullard and Mitra (2002), Marcet and Nicolini (2003), Guesnerie (2009), Branch and McGough (2009, 2010), Woodford (2010), De Grauwe (2011, 2012a,b), De Grauwe and Kaltwasser (2012), Anufriev et al. (2013), Kurz et al. (2013a), Benhabib et al. (2014); see Evans and Honkapohja (2001) and Woodford (2013) for overviews).

³³There are many other central banks with a hierarchical mandate which makes price stability the primary objective for monetary policy, including the central banks of New Zealand, Canada, England, and Sweden.

rational and behavioral versions. In Section 3.3 we first describe the experimental design and the procedures. Then we show the experimental results. Section 3.4 concludes.

3.2 Theory

In this section, we first describe the underlying macroeconomic model. Then we introduce the behavioral model of expectation formation. After that, we compare the outcomes of both models and describe the economic intuition behind these outcomes.

3.2.1 Macroeconomic Model

The economic model we use can be described by the following aggregate New Keynesian equations:

$$(3.1) \quad y_t = \bar{y}_{t+1}^e - \varphi(i_t - \bar{\pi}_{t+1}^e) + g_t$$

$$(3.2) \quad \pi_t = \lambda y_t + \rho \bar{\pi}_{t+1}^e + u_t$$

$$(3.3) \quad i_t = \text{Max}\{\bar{\pi} + \phi_\pi(\pi_t - \bar{\pi}) + \phi_y(y_t - \bar{y}), 0\},$$

where y_t and \bar{y}_{t+1}^e are respectively the actual and average expected output gap, i_t is the nominal interest rate, π_t and $\bar{\pi}_{t+1}^e$ are respectively the actual and average expected inflation rates, g_t and u_t are exogenous disturbances and φ , λ , ρ , ϕ_π and ϕ_y are positive parameters. Equation (3.1) is the aggregate demand equation in which the output gap y_t depends on the average expected future output gap \bar{y}_{t+1}^e and on the real interest rate $i_t - \bar{\pi}_{t+1}^e$. Equation (3.2) is the New Keynesian Phillips curve according to which the inflation rate depends on the output gap and on average expected future inflation. Equation (3.3) is the monetary policy rule implemented by the central bank describing how it reacts to deviations from the inflation target $\bar{\pi}$ and to deviations from the corresponding equilibrium level of the output gap $\bar{y} \equiv (1 - \rho)\bar{\pi}/\lambda$. The coefficients ϕ_π and ϕ_y measure how much the central bank adjusts the nominal interest rate i_t in response to deviations of the inflation rate from its target and of the output gap from its equilibrium level. As usual, the interest rate rule is subject to the zero lower bound, i.e. $i_t \geq 0$. When the zero lower bound is not binding, model (3.1)–(3.3) can be rewritten in matrix form as

$$(3.4) \quad \begin{bmatrix} y_t \\ \pi_t \end{bmatrix} = \Omega \begin{bmatrix} \varphi \bar{\pi}(\phi_\pi - 1) + \varphi \phi_y \bar{y} \\ \lambda \varphi \bar{\pi}(\phi_\pi - 1) + \lambda \varphi \phi_y \bar{y} \end{bmatrix} + \Omega \begin{bmatrix} 1 & \varphi(1 - \phi_\pi \rho) \\ \lambda & \lambda \varphi + \rho + \rho \varphi \phi_y \end{bmatrix} \begin{bmatrix} \bar{y}_{t+1}^e \\ \bar{\pi}_{t+1}^e \end{bmatrix} + \Omega \begin{bmatrix} 1 & -\varphi \phi_\pi \\ \lambda & 1 + \varphi \phi_y \end{bmatrix} \begin{bmatrix} g_t \\ u_t \end{bmatrix},$$

where $\Omega \equiv 1/(1 + \lambda \varphi \phi_\pi + \varphi \phi_y)$.

The economic model described by the aggregate equations (3.1)–(3.3), or equivalently by (3.4), has microfoundations under both rational expectations and under behavioral expectations.³⁴ In the following we will only make use of the aggregate equations presented here.

3.2.2 A Behavioral Model of Expectation Formation

Models with rational expectations are based on the assumption that agents have perfect information and a full understanding of the true model underlying the economy. There is, however, a large body of empirical literature documenting departures from this assumption and showing that agents use heuristics to make forecasts of future (macroeconomic) variables; this behavior is not necessarily a consequence of agents' irrationality but it can also be a "rational" response of agents who face cognitive limitations and have imperfect understanding of the true model underlying the economy (see e.g. Gigerenzer and Todd, 1999, or Gigerenzer and Selten, 2002). Next, we introduce a behavioral model of expectation formation for such an environment.

Let \mathcal{H} denote a set of H different heuristics used by agents to make forecasts. A generic forecasting heuristic $h \in \mathcal{H}$ based on available information at time t can be described as

$$(3.5) \quad x_{h,t+1}^e = f_h(x_{t-1}, x_{t-2} \dots; x_{h,t}^e, x_{h,t-1}^e \dots).$$

In this chapter x is either inflation π or the output gap y . Although agents might use simple rules to predict future inflation and output gap, we impose a certain discipline in the selection of such rules in order to avoid completely irrational behavior. In particular, we introduce a selection mechanism that disciplines the choice of heuristics by agents according to a fitness criterion. This allows agents to learn from past mistakes (the willingness to learn from past mistakes has been called "the most fundamental definition of rational behaviour"; De Grauwe, 2012b). We denote by U_h the fitness measure of a certain forecasting strategy h defined as

$$(3.6) \quad U_{h,t-1} = F(x_{h,t-1}^e - x_{t-1}) + \eta U_{h,t-2},$$

³⁴Under the assumption of a representative agent holding rational expectations, this model represents the standard New Keynesian model discussed for example in Woodford (2003) and Galí (2008). Micro-founded New Keynesian models consistent with heterogeneous expectations have been derived by Branch and McGough (2009), Kurz (2011), Kurz et al. (2013a) and Massaro (2013). System (3.1)–(3.3) corresponds to the model developed by Branch and McGough (2009) augmented with demand and supply shocks, or to the model derived in Kurz (2011) and Kurz et al. (2013a) in which the error terms are interpreted as the deviation of the average of agents' forecasts of their individual future consumption from the average forecast of aggregate consumption and as a similar deviation of price forecasts.

where F is a generic function of the forecast error of heuristic h , and $0 \leq \eta \leq 1$ is a memory parameter, measuring the relative weight agents give to past errors of heuristic h . Performance is completely determined by the most recent forecasting error if $\eta = 0$, while performance depends on all past prediction errors with exponentially declining weights if $0 < \eta < 1$ or with equal weights if $\eta = 1$. If all agents simultaneously update the forecasting rule they use, the fraction of agents choosing rule h in each period t can be modeled as

$$(3.7) \quad n_{h,t} = \frac{\exp(\beta U_{h,t-1})}{\sum_{h=1}^H \exp(\beta U_{h,t-1})}.$$

The multinomial logit expression described in Equation (3.7) can be derived directly from a random utility model (see Manski and McFadden, 1981, and Brock and Hommes, 1997). The parameter $\beta \geq 0$, referred to as “intensity of choice”, reflects the sensitivity of agents to selecting the optimal prediction strategy according to the fitness measure U_h .³⁵ If $\beta = 0$, $n_{h,t}$ is constant for all h , meaning that agents do not exhibit any willingness to learn from past performance; if $\beta = \infty$ all agents adopt the best performing heuristic with probability one. The reinforcement learning model in Equation (3.7) has been extended by Hommes et al. (2005a) and Diks and van der Weide (2005) to include asynchronous updating in order to allow for the possibility that not all agents update their rule in every period (consistent with empirical evidence; see Hommes et al., 2005b, and Anufriev and Hommes, 2012). This yields a generalized version of Equation (3.7) described by

$$(3.8) \quad n_{h,t} = \delta n_{h,t-1} + (1 - \delta) \frac{\exp(\beta U_{h,t-1})}{\sum_{h=1}^H \exp(\beta U_{h,t-1})}.$$

The parameter $0 \leq \delta \leq 1$ introduces persistence in the adoption of forecasting strategies and can be interpreted as the average fraction of individuals who, in each period, stick to their previous strategy.

In order to use this behavioral model for policy analyses or predictions, specific assumptions have to be made on the nature of agents’ forecasting heuristics (in general, the set \mathcal{H} may contain an arbitrary number of forecasting rules). We restrict our attention to a set of four heuristics described in Table 3.1.

The choice of this specific set of heuristics is motivated on empirical grounds. These heuristics were obtained and estimated as descriptions of typical individual forecasting behavior observed in Hommes et al. (2005b), Hommes et al. (2008), and Assenza et al. (2014b) building upon a rich literature on expectation formation (see Hommes, 2011, for a recent sur-

³⁵Equation (3.7) can also be derived from an optimisation problem under rational inattention (see Matějka and McKay, 2015). In this context the parameter β is inversely related to the “shadow cost of information”.

Table 3.1: Set of heuristics

ADA	adaptive rule	$x_{1,t+1}^e = 0.65x_{t-1} + 0.35x_{1,t}^e$
WTR	weak trend-following rule	$x_{2,t+1}^e = x_{t-1} + 0.4(x_{t-1} - x_{t-2})$
STR	strong trend-following rule	$x_{3,t+1}^e = x_{t-1} + 1.3(x_{t-1} - x_{t-2})$
LAA	anchoring and adjustment rule	$x_{4,t+1}^e = 0.5(x_{t-1}^{av} + x_{t-1}) + (x_{t-1} - x_{t-2})$

Notes: x_{t-1}^{av} denotes the average of all observations up to time $t - 1$.

vey).³⁶ Based upon the calibration in these papers, we use the parameters $\beta = 0.4$, $\delta = 0.9$, and $\eta = 0.7$.³⁷

3.2.3 Monetary Policy, Inflation, and Output Gap

A result derived from Model (3.4) under rational expectations is that a policy trade-off is observed between the volatility of the output gap and the volatility of inflation. A decline in output gap volatility resulting from a more active output stabilization policy comes at the price of an increase in inflation volatility (it is reasonable to focus on volatility as for the rational and the behavioral model alike inflation and the output gap are on average at their target, respectively steady state level, for reasonable values of ϕ_π and ϕ_y). This policy trade-off is described in Figure 3.1a, where we show the effect of the parameter ϕ_y (with which the central bank reacts to deviations of the output gap from its steady state level) on inflation volatility. Higher output stabilization, i.e., an increase in the reaction coefficient ϕ_y , comes at the price of higher inflation volatility. The immediate policy implication for a central bank whose main objective is price stability is that it is optimal to set $\phi_y = 0$, i.e. not to react to output gap fluctuations at all (cf. Galí, 2008, and Woodford, 2003).

For the simulations of this graph, the parameter ϕ_π is equal to 1.5 (using different parameters of ϕ_π leads to similar results, which can be seen in Appendix 3.A) and the structural parameters in Equations (3.1)-(3.3) are as estimated in Clarida et al. (2000).³⁸ The inflation target used for the simulations is $\bar{\pi} = 3.5$ (this is the same target that will be used in the experiment, Footnote 43 provides a rationale for this value; the simulations yield similar re-

³⁶The reinforcement learning model described in Equation (3.8) including the set of heuristics presented in Table 3.1 is successfully used in Anufriev and Hommes (2012) to explain different price patterns observed in asset pricing experiments. In Assenza et al. (2014b) the model is used to explain the observed patterns of inflation and output gap in a learning-to-forecast experiment framed in a New Keynesian model similar to (3.4).

³⁷Furthermore, we use the fitness measure $U_h = 100/(1 + |x_h^e - x|)$, which is the function used to incentivize subjects in the experiment described in Section 3.3 (this incentive structure is also used in Adam, 2007, Pfajfar and Zakelj, 2014, and Assenza et al., 2014b, among others). The simulation results in Section 3.2.3 are qualitatively robust to alternative specifications of the fitness metric, such as using a quadratic function.

³⁸Thus, $\rho = 0.99$, $\lambda = 0.3$, and $\varphi = 1$ (for quarterly data). The shocks g_t and u_t are independent and normally distributed with standard deviation 0.1. The number of simulations for each value of ϕ_y is 10000.

sults for different values of $\bar{\pi}$). This inflation target leads to a steady state level of the output gap of $\bar{y} = 0.1166667$. Inflation volatility is measured by $v(\pi) = \frac{1}{T-1} \sum_{t=2}^T (\pi_t - \pi_{t-1})^2$, with T denoting the total number of periods. This measure has some properties that make it preferable to other measures of volatility (using alternative measures yields similar results).³⁹

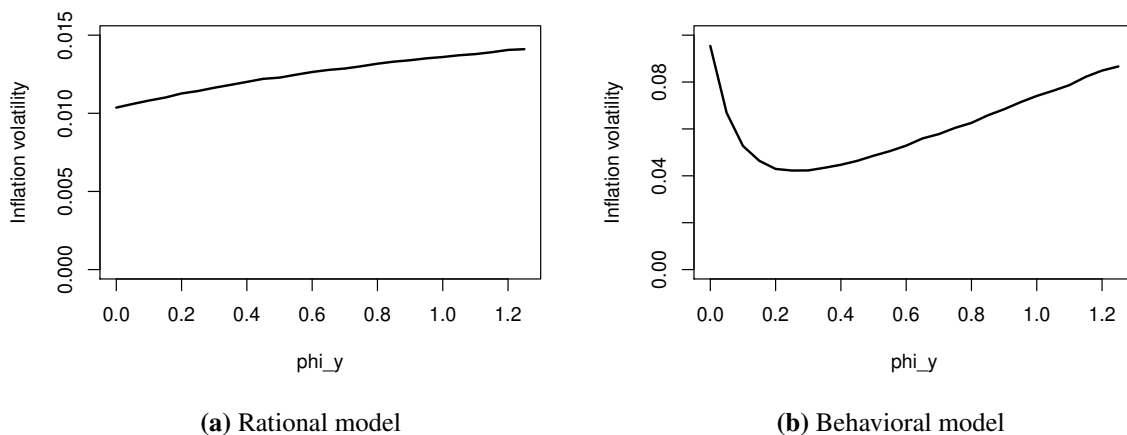


Figure 3.1: Inflation volatility as a function of ϕ_y for the rational and the behavioral model

Notes: This figure shows the effect of the output gap reaction coefficient on inflation volatility.

In Figure 3.1b, we show the effect of the parameter ϕ_y on inflation volatility when expectations are formed according to the behavioral model described in Section 3.2.2 (note that the scales in Figures 3.1a and 3.1b are different; the overall level of inflation volatility is higher under behavioral expectations than under rational expectations). In contrast to the simulation results under rational expectations, the graph of inflation volatility as a function of ϕ_y has a *U*-shape.⁴⁰ Thus, starting from $\phi_y = 0$, the central bank can simultaneously decrease volatility of inflation and output gap by also reacting with its monetary policy to deviations of the output gap from its steady state level (Figure 3.2 depicts output gap volatility as a function of ϕ_y ; as ϕ_y increases output gap volatility decreases under both rational and behavioral expectations). These results are qualitatively robust to changes in starting values, inflation target, and parameters involved in the expectation formation.⁴¹ Hence, under be-

³⁹One could also use $v(\pi) = \frac{1}{T-1} \sum_{t=2}^T |\pi_t - \pi_{t-1}|$ or simply the standard deviation to measure volatility. An advantage of the measure we use when compared to the standard deviation is that short-term fluctuations are accounted for differently. The standard deviation of a time series does not change after a permutation of its values, although this can change how much the series fluctuates (imagine one time series that always alternates between the same level of high and low and one that first stays at the same low value for a while and then switches to the same high value).

⁴⁰The starting values used for the simulations of the behavioral model are $\pi_{start} = 3.0$ and $y_{start} = 0.5$, Appendix 3.A provides graphs for different starting values, which are also U-shaped.

⁴¹Similar results also arise in a different macroeconomic model when employing simplistic behavioral rules

havioral expectations, there is a broader scope for output stabilization.

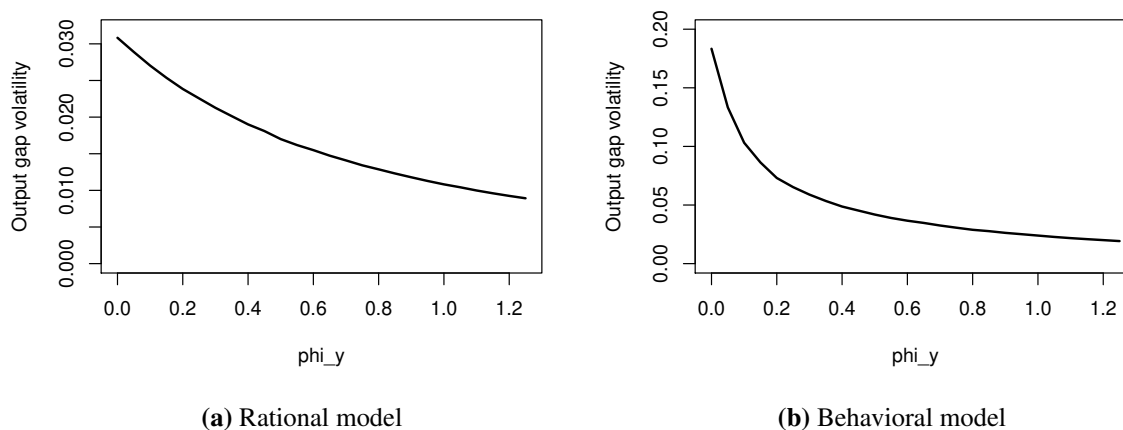


Figure 3.2: Output gap volatility as a function of ϕ_y for the rational and the behavioral model

Notes: This figure shows the effect of the output gap reaction coefficient on output gap volatility.

Now we turn to the intuition of these results. Considering the outcome simulated with rational expectations (Figure 3.1a), it would be easy to fall prey to the following simple, but incorrect, intuition: “If there are two variables, targeting one variable will always come at the expense of the other variable”. This is in general not the case, the intuition is slightly more complex. Homogeneous rational expectations are strictly forward looking and in this model always equal to the inflation target and the corresponding steady state level of the output gap, respectively (assuming that $\phi_\pi + \phi_y(1 - \rho)/\lambda > 1$, which ensures a determinate model solution, see e.g. Woodford, 2003). These expectations do not depend in any way on the current level of inflation and output gap or on any past behavior. It is exactly via the dependence of expectations on (past) actual variables that reacting to the output gap can also pay off in terms of inflation volatility. To illustrate this, imagine inflation and output gap staying constant at $\bar{\pi}$ and \bar{y} and a combination of shocks arriving in one period that would lead (without any reaction by the central bank) to inflation staying constant and the output gap being above the steady state level. Should the central bank react to this shock if it only cares about inflation? The rational expectations answer would be no; inflation is at its target and in the next period one would (assuming no further shocks) again be at the inflation target and the steady state level of the output gap, because expectations do not react to the past. However, under behavioral expectations, what happens today matters for the future. If there

of expectation formation (De Grauwe, 2011, 2012a). Such a non-monotonic trade-off between inflation and output gap volatility can also arise in sticky information economies in which the degree of attentiveness or the rate at which agents update their information is endogenized (Branch et al., 2009).

is some trend-following behavior, a higher output gap now will lead agents to revise their expectations of the future output gap upwards, which will in turn lead to a higher realized output gap in the future which will lead to upward pressure on inflation. Therefore it can be beneficial for the central bank to curb the increase of the output gap now (at the expense of slightly lower inflation now) in order to dampen the upward pressure on inflation in the future. However, if the monetary authority puts too much weight on output gap stabilization, the ensuing fluctuations in inflation dominate the stabilization bonus provided by less volatile output, leading to higher inflation volatility.

Regarding the results from the behavioral model, one can also look at the heuristics involved. Of the four heuristics, only one heuristic is stabilizing, namely the adaptive rule (ADA). All other heuristics have some component extrapolating trends (coordination on trend-following heuristics is generally associated with high volatility, see e.g. Anufriev and Hommes, 2012, Bao et al., 2012, Assenza et al., 2014b, and Pfajfar and Zakelj, 2014). Thus, the volatility of a variable tends to be lower when the adaptive rule performs relatively well. Therefore, values of ϕ_y that lead to a relatively large fraction of the adaptive rule for inflation forecasting can be expected to also lead to low inflation volatility. This is indeed the case and can be seen in Figure 3.3, where the average fractions of heuristics used for inflation forecasting and output gap forecasting are shown as a function of ϕ_y . For inflation forecasting, starting from $\phi_y = 0$, increasing the reaction to the output gap first increases the fraction of agents using the adaptive rule, but after some level of ϕ_y , increasing it further reduces the fraction of agents using the adaptive rule. For output gap forecasting, it is not surprising to see that the fraction of agents using the adaptive rule increases monotonically with ϕ_y .

3.3 Experiment

The only task for subjects in the experiment is to forecast inflation and the output gap. These forecasts are then used to calculate subsequent realizations. The model underlying the experimental economy is the macroeconomic model described in Section 3.2.1 (with the same calibration of macroeconomic parameters as before). Before we describe the experiment in more detail, we now explain the treatments and hypotheses. The design of the experiment and the hypotheses can be motivated with the theory described in Section 3.2; however, the experiment is also informative without this theory, as it can be seen as a mere investigation of the effects of a change in monetary policy in a controlled laboratory environment.

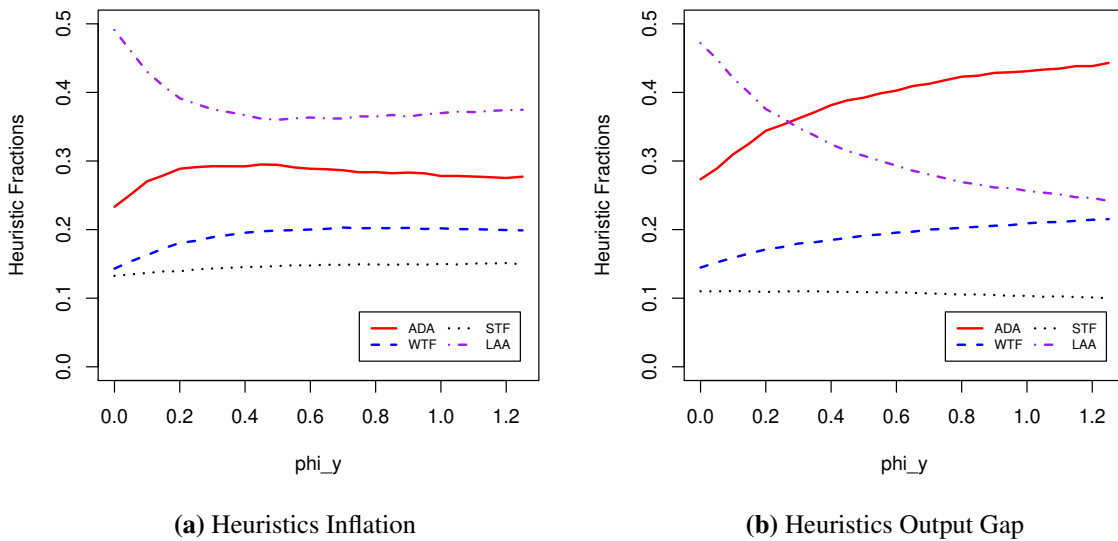


Figure 3.3: Fractions of heuristics used for behavioral expectations of output gap and inflation

Notes: This figure shows the average fractions of heuristics used as functions of the output gap reaction coefficient.

3.3.1 Treatments and Hypotheses

There are two treatments, $T1$ (“inflation targeting only”) and $T2$ (“inflation and output gap targeting”). The only difference between the treatments lies in the Taylor rule describing monetary policy. In $T1$, the parameters of the Taylor rule are $\phi_\pi = 1.5$ and $\phi_y = 0$, whereas they are $\phi_\pi = 1.5$ and $\phi_y = 0.5$ in $T2$. That is, the only difference between the treatments is that in $T1$ the central bank only targets inflation whereas it also targets the output gap on top of inflation in $T2$.

We are interested in testing the null-hypothesis (which can be derived from the rational expectations model in Section 3.2) that inflation volatility in $T1$ is less or equal to inflation volatility in $T2$ against the alternative hypothesis (which can be derived from the behavioral model) that inflation volatility is greater in $T1$ than in $T2$. Figure 3.4 summarizes these hypotheses, i.e. the treatment effects one can expect arising from rational expectations and from the behavioral model described.⁴²

The inflation and output gap expectations arising from a continuum of rational agents are $\bar{\pi}^e = \bar{\pi}$ and $\bar{y}^e = \bar{y}$. In the experiment, the number of subjects per experimental economy is six. Evidence from other experiments indicates that four to six subjects are enough to

⁴²While some people may argue that the best test of the models is to compare subjects’ forecasts to the model predictions (in which the behavioral model does much better), others could question such a comparison on the ground that it is a within-treatment comparison; the directionally different hypotheses in our experiment make it a cleaner test (in laboratory experiments, the comparative statics of treatment comparisons are generally considered to be most robust and relevant; see Schram, 2005, Falk and Heckman, 2009, and Charness and Kuhn, 2011).

	T1 ($\phi_\pi = 1.5, \phi_y = 0$)	T2 ($\phi_\pi = 1.5, \phi_y = 0.5$)
Null-hypothesis (rational exp.)		→
Alternative hypothesis (behavioral exp.)		→

Figure 3.4: Hypotheses about inflation volatility

justify the use of the competitive equilibrium as equilibrium concept (see e.g. Huck et al., 2004). Note, however, that also in a game theoretic analysis the unique Nash equilibrium is forecasting $\bar{\pi}$ and \bar{y} .

3.3.2 Course of Events and Implementation

The design is a between subject design with within session randomization. In the beginning, all participants are divided into groups (experimental economies) of six. Subjects only interact with other subjects in their group, without knowing who they are. The task subjects have is to make two-period-ahead forecasts of inflation and output gap. The average forecasts of all subjects in one group are then used to calculate the realizations of inflation and output gap according to model equations (3.1)–(3.3) (only the average forecasts $\bar{\pi}_{t+1}^e$ and \bar{y}_{t+1}^e are needed to calculate the realizations π_t and y_t). The inflation target of the central bank in the experiment is $\bar{\pi} = 3.5$.⁴³ When making their forecasts for period $t + 1$, the information subjects can see on their screen (as numbers and partly also in graphs) is the following: all realizations of inflation, output gap, and interest rate up to period $t - 1$, their own forecasts of inflation and output gap up to period t and their scores stating how close their past forecasts were to realized values up to period $t - 1$ (these scores determine the payments). Figure 3.5 shows a screenshot of the experiment (a larger version of the same screenshot can be found in Appendix 3.C).

Subjects' payments depend on their forecasting performance. At the end of the experiment it is determined randomly for each participant whether she is paid for inflation forecasting or output gap forecasting. The total scores for inflation and output gap forecasting are the sums of the respective forecasting scores over all periods. This score is for subject

⁴³This number has been chosen so as to be (i) large enough to have some distance from the zero lower bound as this is not supposed to be a liquidity trap experiment, (ii) different from focal points such as 2% or 2.5%, which are standard inflation targets in the real world, so that we can observe some learning in the experiment, and (iii) low enough so as not to be too far away from zero not to make the approximation from the log-linearized equations too imprecise.

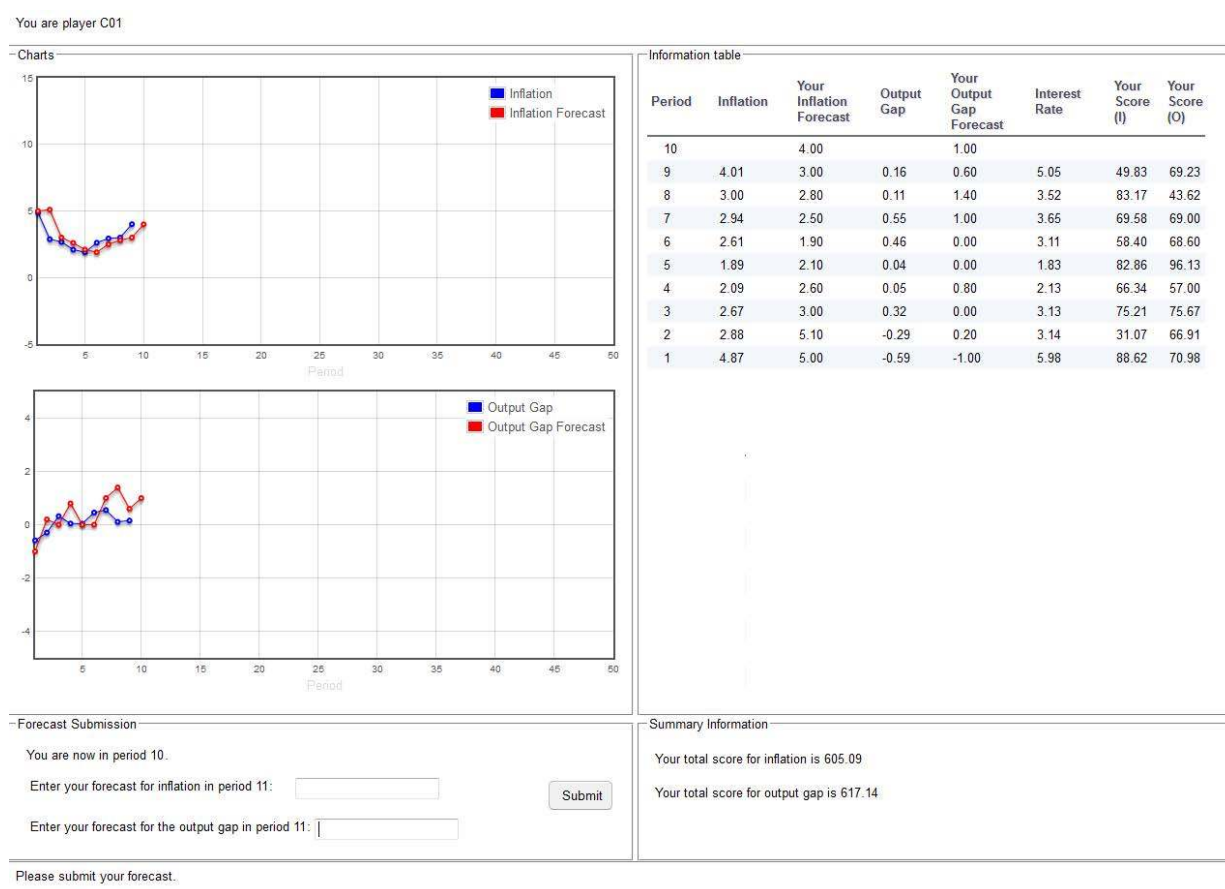


Figure 3.5: Screenshot

i 's inflation forecast in period t equal to $100/(1 + |\pi_{t,i}^e - \pi_t|)$, where $\pi_{t,i}^e$ denotes subject i 's forecast for period t and π_t the realized value of this period. The score for output gap forecasting is calculated analogously. This means that subjects' payments decrease with the distance of the realizations from their forecasts.

In the instructions, subjects receive a qualitative description of the economy, describing the mechanisms governing the model equations. Concerning monetary policy, subjects in both treatments are only told that the central bank decreases the interest rate if it wants to increase inflation or output gap and that it increases the interest rate if it wants to decrease inflation or output gap.⁴⁴ Except for the precise formulation of the equations of the macroeconomic model, the instructions contain full information about the experiment (i.e. on the number of subjects per group, payments, etc.). The complete instructions can be found in Appendix 3.B.

The experiment was programmed in java and conducted at the CREED laboratory at the

⁴⁴As the experiment uses two-period ahead forecasts, subjects are asked after having finished the instructions to enter forecasts for periods 1 and 2 simultaneously. Subjects therefore receive some indication of reasonable values by being told in the instructions that in economies similar to the one at hand inflation has historically been between -5% and 10% and the output gap between -5% and 5% .

University of Amsterdam. The experiment was conducted with 192 subjects recruited from the CREED subject pool (32 groups of six subjects each, distributed over nine sessions). After each session, participants had to fill out a short questionnaire. Participants were primarily undergraduate students, the average age was slightly above 22 years. About half of the participants were female, about two thirds were majoring in economics or business, and about two thirds were Dutch. During the experiment, ‘points’ were used as currency. These points were exchanged for euros at the end of each session at an exchange rate of 0.75 euros per 100 points. The experiment lasted around two hours, and participants earned on average 30.45 euros. The series of error terms used in the model equations (g_t and u_t in equations 3.1 and 3.2) were different from group to group within each treatment, but the sets of noise series in both treatments were the same.⁴⁵

3.3.3 Results

There are data of 32 different groups, 17 in $T1$ and 15 in $T2$. The groups’ actions do not influence one another in any way, thus the observations at the group level can be treated as statistically independent. To get a good overview of the data, consider Figures 3.6 and 3.7 (the data for each group including all individual forecasts can be found in Appendix 3.C).

Figure 3.6 gives an overview of inflation in all experimental economies, separately for $T1$ and $T2$. Each line corresponds to the inflation in one experimental economy, tracked over all 50 periods of the experiment. Almost all economies are close to the inflation target after 50 periods and for the economies with inflation still oscillating around the target the amplitude of these oscillations is decreasing. That many economies are converging to the steady state over the course of the experiment is not necessarily surprising as there are 50 periods without any changes to the underlying model (cf. Pfajfar and Zakelj, 2014, and Assenza et al., 2014b). The figure shows that inflation is indeed less volatile in $T2$ when also the output gap is targeted than in $T1$, as predicted by the behavioral model (consider for example inflation after half of the periods: Except for one economy, all economies in $T2$ already exhibit relatively

⁴⁵ Before conducting the experiment, two pilot sessions were conducted (with a total of six groups). The pilot sessions differ from the actual experiment as follows: the error terms added to the model equations had a larger standard deviation, a different inflation target was used, and subjects in the pilot did not receive any information on the number of participants in each group. For two of the groups also a different combination of parameters for the Taylor rule was used.

We exclude two of the groups from the analysis (including these two groups, the experiment was conducted with 204 subjects). One of the groups was excluded, because of a very large typo (30 instead of 3.0; the corresponding participant notified us about this typo in the post-experiment questionnaire). The other group was excluded due to severe misunderstandings of one subject, who systematically stayed very far from the actual realizations (thereby also losing a lot of money). Our conclusions do not change if we include these groups in our analysis. The realizations and forecasts of inflation and output gap for these two groups are shown in Figure 3.20, Appendix 3.C.

low volatility and are close to the target, while in $T1$ many economies still exhibit wildly fluctuating inflation).

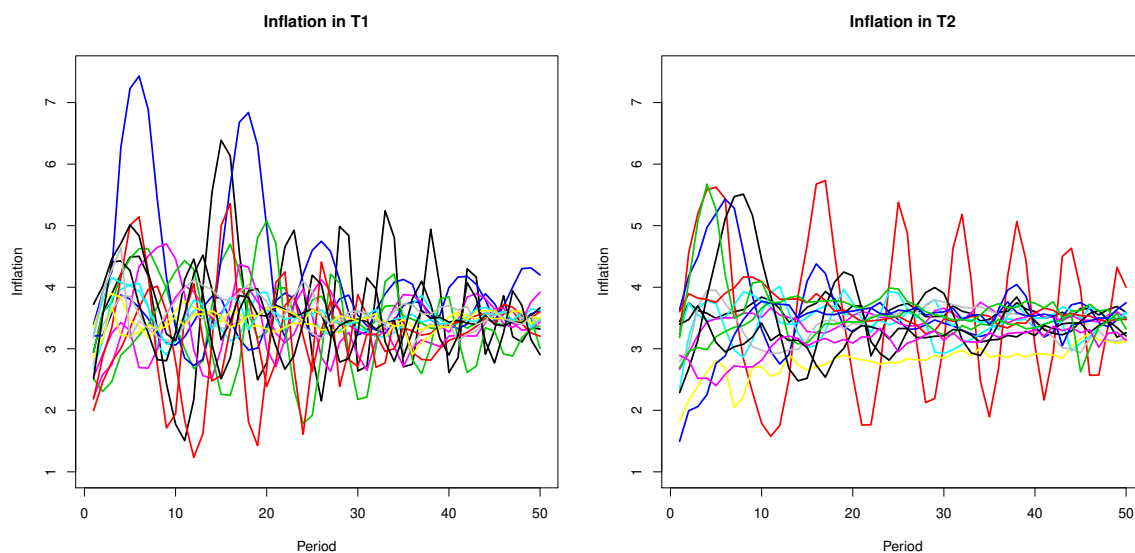


Figure 3.6: Realized inflation for all groups in both treatments

Notes: Each line represents realized inflation in one economy. On the horizontal axis is the number of periods (1 to 50), on the vertical axis inflation in percent (from 1 to 7.5).

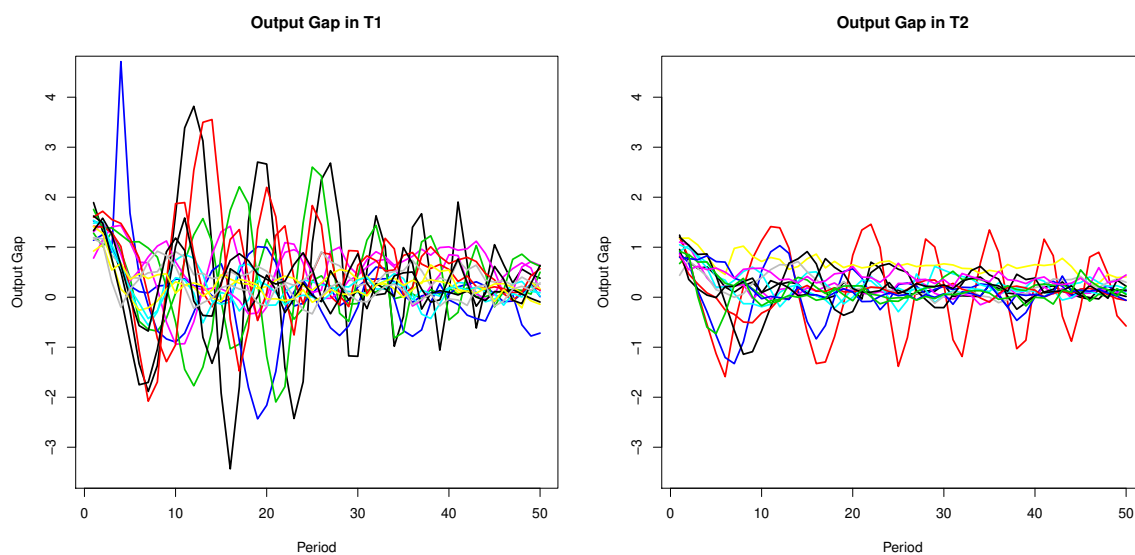


Figure 3.7: Realized output gap for all groups in both treatments

Notes: Each line represents realized output gap in one economy. On the horizontal axis is the number of periods (1 to 50), on the vertical axis output gap in percent (from -3.5 to 4.5). Each line has the same color as the line for the same group's inflation in Figure 3.6.

Figure 3.7 shows the output gap for all experimental economies. Here, the differences are even larger, the output gap is much more volatile in $T1$ than in $T2$. This was to be expected, both models predict that the output gap is more stable when it is also targeted by the central bank. While inflation and output gap volatility are quite different between the treatments, these variables generally fluctuate around their steady state values: The mean of inflation over all 50 periods is between 3.13 and 4.33 in $T1$ and between 2.79 and 3.76 in $T2$ for all groups, the mean of the output gap is between -0.12 and 0.70 in $T1$ and between 0.05 and 0.66 in $T2$.

We now turn to more detail about inflation volatility in the experiment. As in Section 3.2.3, we use $v(\pi) = \frac{1}{T-1} \sum_{t=2}^T (\pi_t - \pi_{t-1})^2$ as measure of inflation volatility (see Footnote 39). The values of this measure in all experimental economies can be seen in Figure 3.8 where the empirical cumulative distribution functions (ECDFs) are drawn, for groups in both treatments (for each value on the horizontal axis, the ECDF shows on the vertical axis the fraction of groups in each treatment with inflation volatility less or equal to this value; the dots stand for the actual observations). It can easily be seen that inflation volatility is lower in $T2$ than in $T1$. In fact, the whole ECDF of observations in $T2$ lies to the left of the ECDF of observations in $T1$ (the single one high value in $T2$ corresponds to the oscillating red line in the right graph of Figure 3.6).

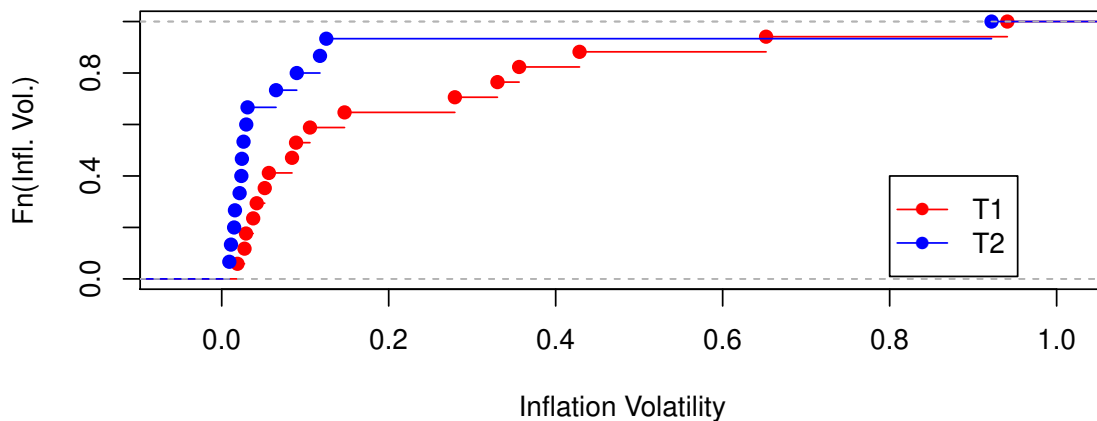


Figure 3.8: Empirical distribution functions of inflation volatility

Notes: For each value on the horizontal axis, the fraction of observations with inflation volatility less or equal to this value (i.e. the ECDF) is shown on the vertical axis, separately for $T1$ and $T2$.

In order to test the statistical significance of this finding we use a Wilcoxon rank-sum test. We test the null-hypothesis that inflation volatility is lower or equal in $T1$ than in $T2$ against the alternative hypothesis that inflation volatility is lower in $T2$.⁴⁶ This test rejects the

⁴⁶Strictly speaking, the Wilcoxon rank-sum test tests the null-hypothesis that the distribution shifts to the right (from $T1$ to $T2$) or that it does not change.

null-hypothesis ($p = 0.006$). The Wilcoxon rank-sum test has the advantage that it makes very unrestrictive assumptions on the underlying data. Note, however, that the results are robust to employing different tests.⁴⁷

3.4 Concluding Remarks

We have conducted a learning-to-forecast experiment to test the predictions of a macroeconomic model with behavioral expectations. This behavioral model yields results that are in contrast to the results from the same macroeconomic model based on rational expectations. Namely, the behavioral model yields that inflation volatility can be reduced if the central bank reacts to the output gap on top of inflation. The predictions of the behavioral model are supported by the outcomes of our learning-to-forecast experiment in which only the monetary policy reaction function of the central bank was modified as a treatment variable.

These results are relevant for monetary policy analysis and important for central banks. They give support to a trade-off between inflation and output-gap that is different than usually assumed based on the standard models with rational expectations. The policy implications are particularly straightforward for central banks only aiming at price stability, such as for example the ECB; these banks should react to the output gap even if they are ultimately only interested in price stability.

⁴⁷The data are not normally distributed, but the logarithms of the data look rather close to a normal distribution (and are statistically not significantly different from it according to a Kolmogorov-Smirnov test). A t-test on the logarithms of the data also rejects the null-hypothesis ($p = 0.009$).

Appendix 3.A Additional Graphs from Simulations of the Macroeconomic Model

Figure 3.9 shows inflation volatility as a function of the output gap reaction coefficient ϕ_y for the model assuming rational expectations, similarly to Figure 3.1a. The graph now shows multiple coefficients of ϕ_π simultaneously (from top to bottom the lines correspond to ϕ_π -values of 1.4, 1.5, 1.6, and 1.7). Figure 3.10 shows the same graph for the behavioral model (again the lines correspond to ϕ_π -values of 1.4, 1.5, 1.6, and 1.7, from top to bottom).

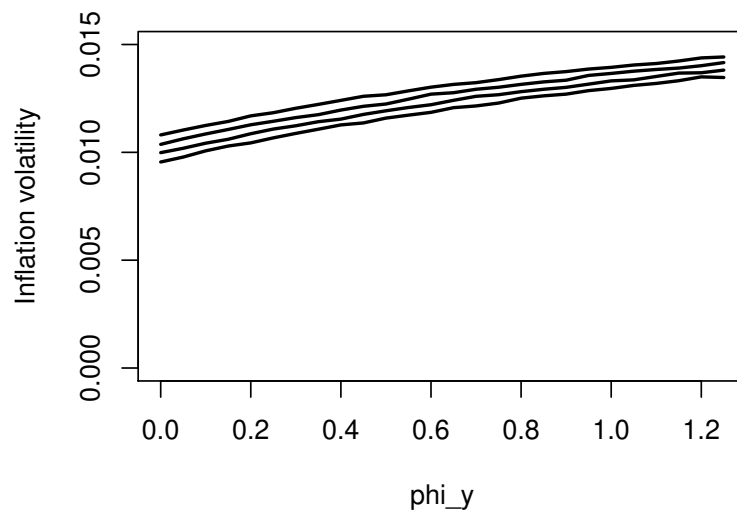


Figure 3.9: Inflation volatility in the rational model as a function of ϕ_y for different values of ϕ_π

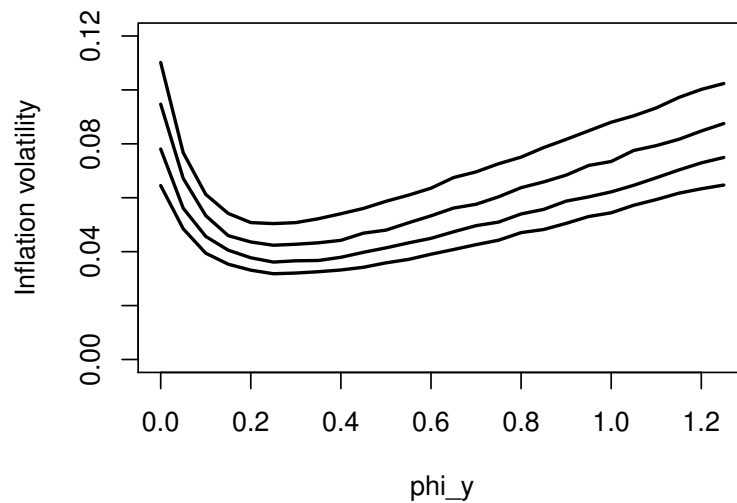
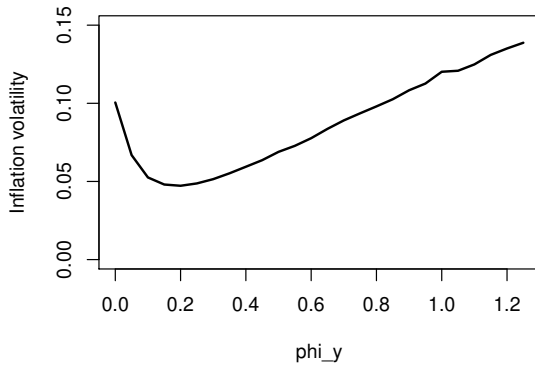


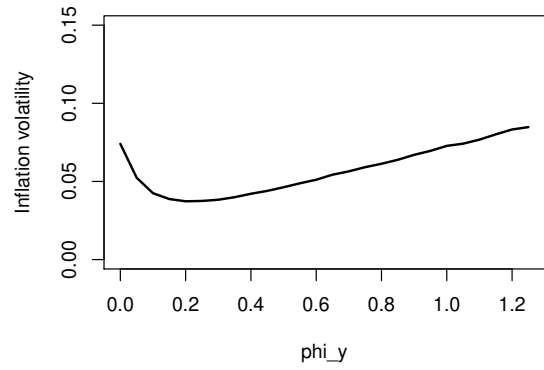
Figure 3.10: Inflation volatility in the behavioral model as a function of ϕ_y for different values of ϕ_π

Figures 3.11 and 3.12 show graphs similar to Figure 3.1b for different combinations of starting values of inflation and output gap (i.e. inflation and output gap are set to these starting values in the first two periods). In all cases the U-shape arises similarly to Figure 3.1b.

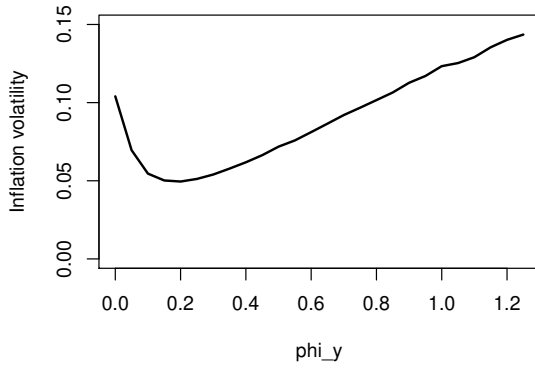
3.A. Additional Graphs from Simulations of the Macroeconomic Model



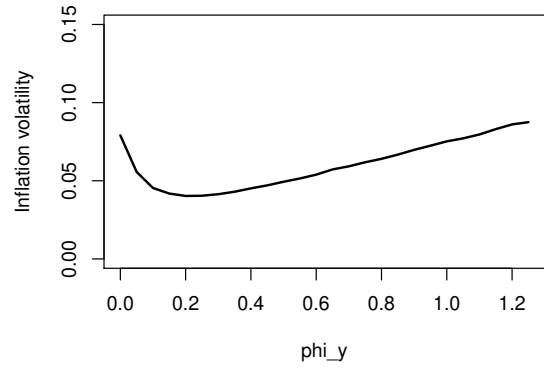
(a) $\pi_{start} = 2.5, y_{start} = -0.5$



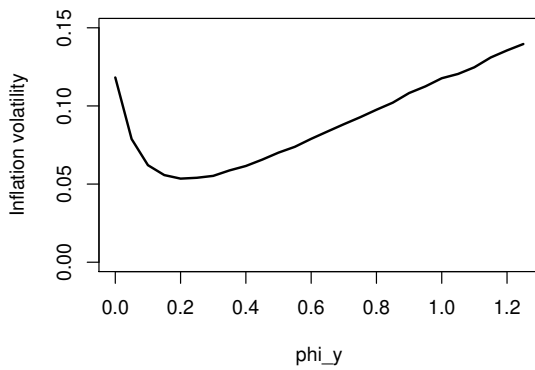
(b) $\pi_{start} = 3.0, y_{start} = -0.5$



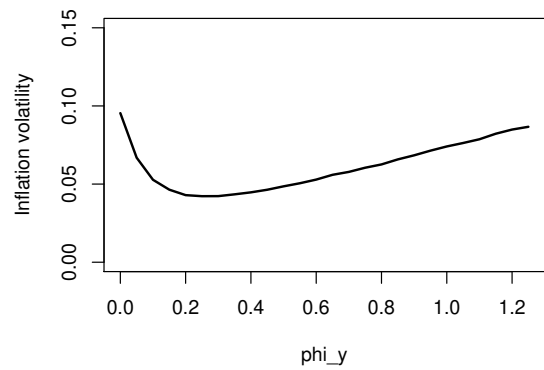
(c) $\pi_{start} = 2.5, y_{start} = 0$



(d) $\pi_{start} = 3.0, y_{start} = 0$



(e) $\pi_{start} = 2.5, y_{start} = 0.5$



(f) $\pi_{start} = 3.0, y_{start} = 0.5$

Figure 3.11: Inflation volatility in the behavioral model for different starting values

Notes: This figure shows the effect of parameter ϕ_y on inflation volatility for different starting values of y and π ($\phi_\pi = 1.5$ throughout).

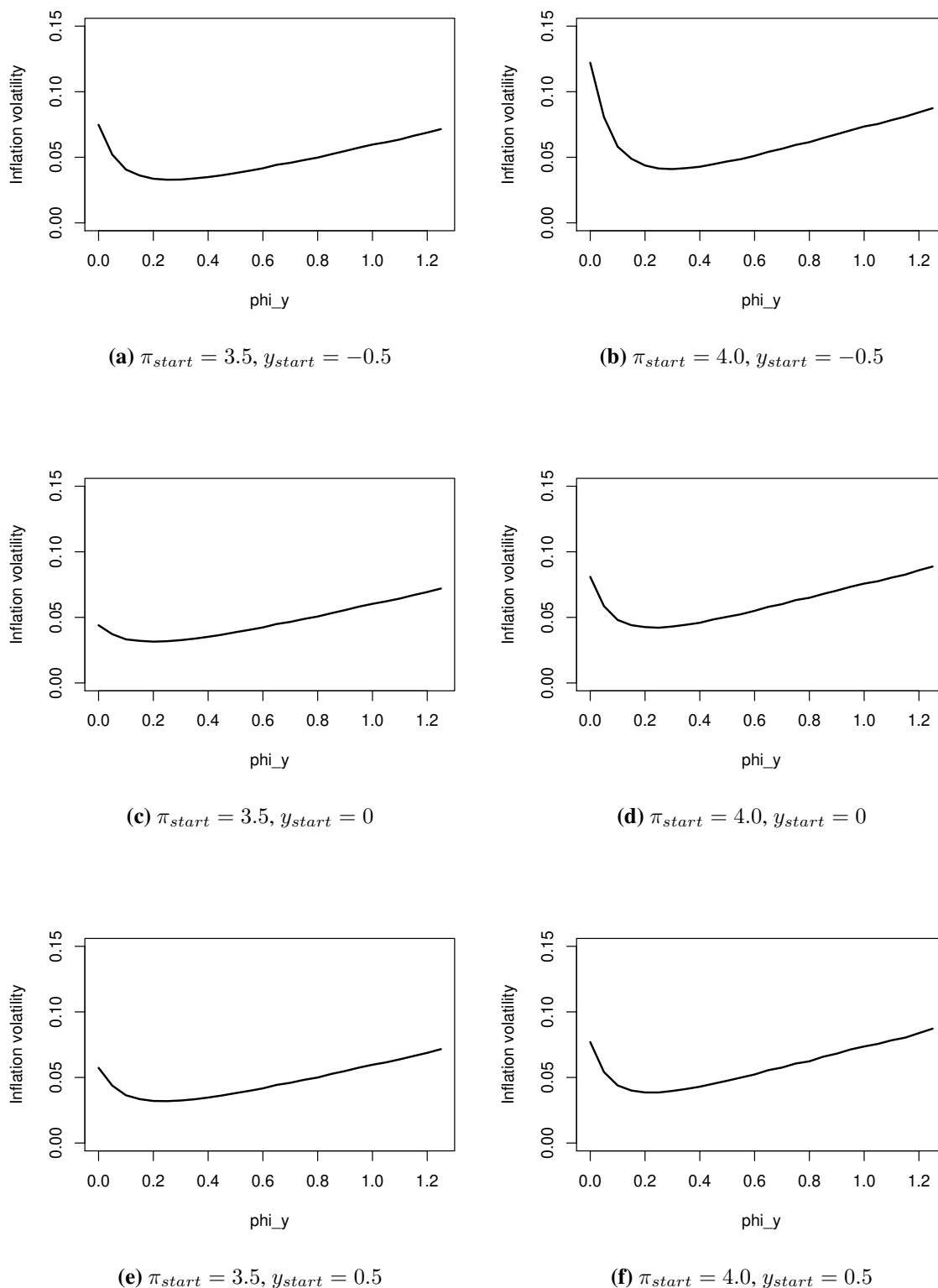


Figure 3.12: Inflation volatility in the behavioral model for different starting values

Notes: This figure shows the effect of parameter ϕ_y on inflation volatility for different starting values of y and π ($\phi_\pi = 1.5$ throughout).

Appendix 3.B Instructions in the Experiment

Subjects in the experiment received the following instructions (as subjects only received qualitative information on the model governing the experimental economy the instructions are the same for both treatments):

Instructions

Welcome to this experiment! The experiment is anonymous, the data from your choices will only be linked to your station ID, not to your name. You will be paid privately at the end, after all participants have finished the experiment. After the main part of the experiment and before the payment you will be asked to fill out a short questionnaire. On your desk you will find a calculator and scratch paper, which you can use during the experiment.

During the experiment you are not allowed to use your mobile phone. You are also not allowed to communicate with other participants. If you have a question at any time, please raise your hand and someone will come to your desk.

General information and experimental economy

All participants will be randomly divided into groups of six people. The group composition will not change during the experiment. You and all other participants will take the roles of statistical research bureaus making predictions of inflation and the so-called “output gap”. The experiment consists of 50 periods in total. In each period you will be asked to predict inflation and output gap for the next period. The economy you are participating in is described by three variables: inflation π_t , output gap y_t and interest rate i_t . The subscript t indicates the period the experiment is in. In total there are 50 periods, so t increases during the experiment from 1 to 50.

Inflation

Inflation measures the percentage change in the price level of the economy. In each period, inflation depends on inflation predictions of the statistical research bureaus in the economy (a group of six participants in this experiment), on actual output gap and on a random term. There is a positive relation between the actual inflation and both inflation predictions and actual output gap. This means for example that if the inflation predictions of the research bureaus increase, then actual inflation will also increase (everything else equal). In economies similar to this one, inflation has historically been between -5% and 10% .

Output gap

The output gap measures the percentage difference between the Gross Domestic Product (GDP) and the natural GDP. The GDP is the value of all goods produced during a period in the economy. The natural GDP is the value the total production would have if prices in the economy were fully flexible. If the output gap is positive (negative), the economy therefore produces more (less) than the natural GDP. In each period the output gap depends on inflation predictions and output gap predictions of the statistical bureaus, on the interest rate and on a random term. There is a positive relation between the output gap and inflation predictions and also between the output gap and output gap predictions. There is a negative relation between the output gap and the interest rate. In economies similar to this one, the output gap has historically been between -5% and 5% .

Interest Rate

The interest rate measures the price of borrowing money and is determined by the central bank. If the central bank wants to increase inflation or output gap it decreases the interest rate, if it wants to decrease inflation or output gap it increases the interest rate.

Prediction task

Your task in each period of the experiment is to predict inflation and output gap in the next period. When the experiment starts, you have to predict inflation and output gap for the first two periods, i.e. π_1^e and π_2^e , and y_1^e and y_2^e . The superscript e indicates that these are predictions. When all participants have made their predictions for the first two periods, the actual inflation (π_1), the actual output gap (y_1) and the interest rate (i_1) for period 1 are announced. Then period 2 of the experiment begins. In period 2 you make inflation and output gap predictions for period 3 (π_3^e and y_3^e). When all participants have made their predictions for period 3, inflation (π_2), output gap (y_2), and interest rate (i_2) for period 2 are announced. This process repeats itself for 50 periods.

Thus, in a certain period t when you make predictions of inflation and output gap in period $t + 1$, the following information is available to you:

- Values of actual inflation, output gap and interest rate up to period $t - 1$;
- Your predictions up to period t ;
- Your prediction scores up to period $t - 1$.

Payments

Your payment will depend on the accuracy of your predictions. You will be paid either for predicting inflation or for predicting the output gap. The accuracy of your predictions is measured by the absolute distance between your prediction and the actual values (this distance is the prediction error). For each period the prediction error is calculated as soon as the actual values are known; you subsequently get a prediction score that decreases as the prediction error increases. The table below gives the relation between the prediction error and the prediction score. The prediction error is calculated in the same way for inflation and output gap.

Prediction error	0	1	2	3	4	9
Score	100	50	33.33	25	20	10

Example: If (for a certain period) you predict an inflation of 2%, and the actual inflation turns out to be 3%, then you make an absolute error of $3\% - 2\% = 1\%$. Therefore you get a prediction score of 50. If you predict an inflation of 1%, and the actual inflation turns out to be negative 2% (i.e. -2%), you make a prediction error of $1\% - (-2\%) = 3\%$. Then you get a prediction score of 25. For a perfect prediction, with a prediction error of zero, you get a prediction score of 100. The figure below shows the relation between your prediction score (vertical axis) and your prediction error (horizontal axis). Points in the graph correspond to the prediction scores in the previous table.

[Figure 3.13 appears here in the experimental instructions.]

At the end of the experiment, you will have two total scores, one for inflation predictions and one for output gap predictions. These total scores simply consist of the sum of all prediction scores you got during the experiment, separately for inflation and output gap predictions. **When the experiment has ended, one of the two total scores will be randomly selected for payment.**

Your final payment will consist of 0.75 euro for each 100 points in the selected total score (200 points therefore equals 1.50 euro). This will be the only payment from this experiment, i.e. you will not receive a show-up fee on top of it.

Computer interface

The computer interface will be mainly self-explanatory. The top right part of the screen will show you all of the information available up to the period that you are in (in period t , i.e. when you are asked to make your prediction for period $t + 1$, this will be actual

inflation, output gap, and interest rate until period $t - 1$, your predictions until period t , and the prediction scores arising from your predictions until period $t - 1$ for both inflation (I) and output gap (O)). The top left part of the screen will show you the information on inflation and output gap in graphs. The axis of a graph shows values in percentage points (i.e. 3 corresponds to 3%). **Note that the values on the vertical axes may change during the experiment and that they are different between the two graphs – the values will be such that it is comfortable for you to read the graphs.**

In the bottom left part of the screen you will be asked to enter your predictions. When submitting your prediction, use a decimal point if necessary (not a comma). For example, if you want to submit a prediction of 2.5% type “2.5”; for a prediction of -1.75% type “-1.75”. The sum of the prediction scores over the different periods are shown in the bottom right of the screen, separately for your inflation and output gap predictions.

At the bottom of the screen there is a status bar telling you when you can enter your predictions and when you have to wait for other participants.

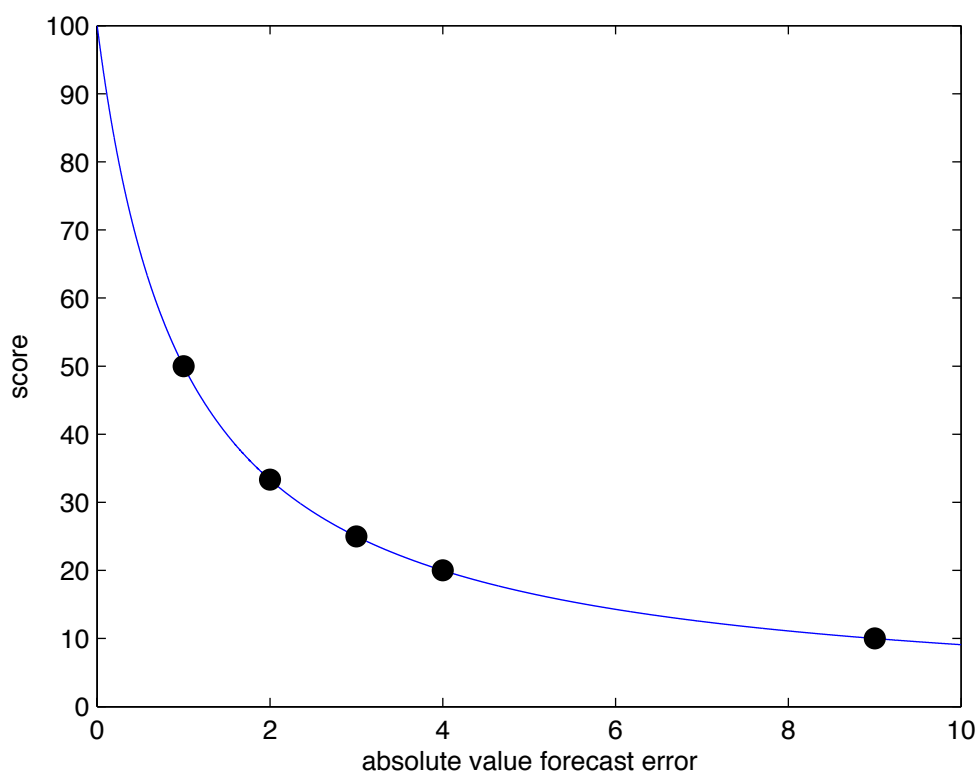


Figure 3.13: Relation score and forecast error (not labeled in the instructions)

Appendix 3.C Additional Graphs of the Experimental Data and Screenshot

Figures 3.14 to 3.20 show the realizations and forecasts of inflation and output gap. Each graph corresponds to one group of six people (one experimental economy). The thick black line shows the realization of inflation, the thin dashed black lines show the inflation forecasts of the six individuals in the group. The thick gray line shows the realization of the output gap and the thin dashed gray lines show the output gap forecasts of all individuals in a group. On the horizontal axis are the periods (from 1 to 50), on the vertical axis are the values of inflation and output gap in percent (the numbers on the vertical axis reach from -3 to 8). The upper horizontal line corresponds to the steady state value of inflation ($\bar{\pi} = 3.5$), the lower horizontal line corresponds to the steady state value of the output gap ($\bar{y} = 0.1166667$). Figures 3.14 to 3.16 show all groups of treatment $T1$, Figures 3.17 to 3.19 show the groups of treatment $T2$. Figure 3.20 shows the two groups (from $T2$) that have been excluded from the analysis as explained in Footnote 45.

Figure 3.21 shows a screenshot (a larger version of the screenshot already used in Figure 3.5).

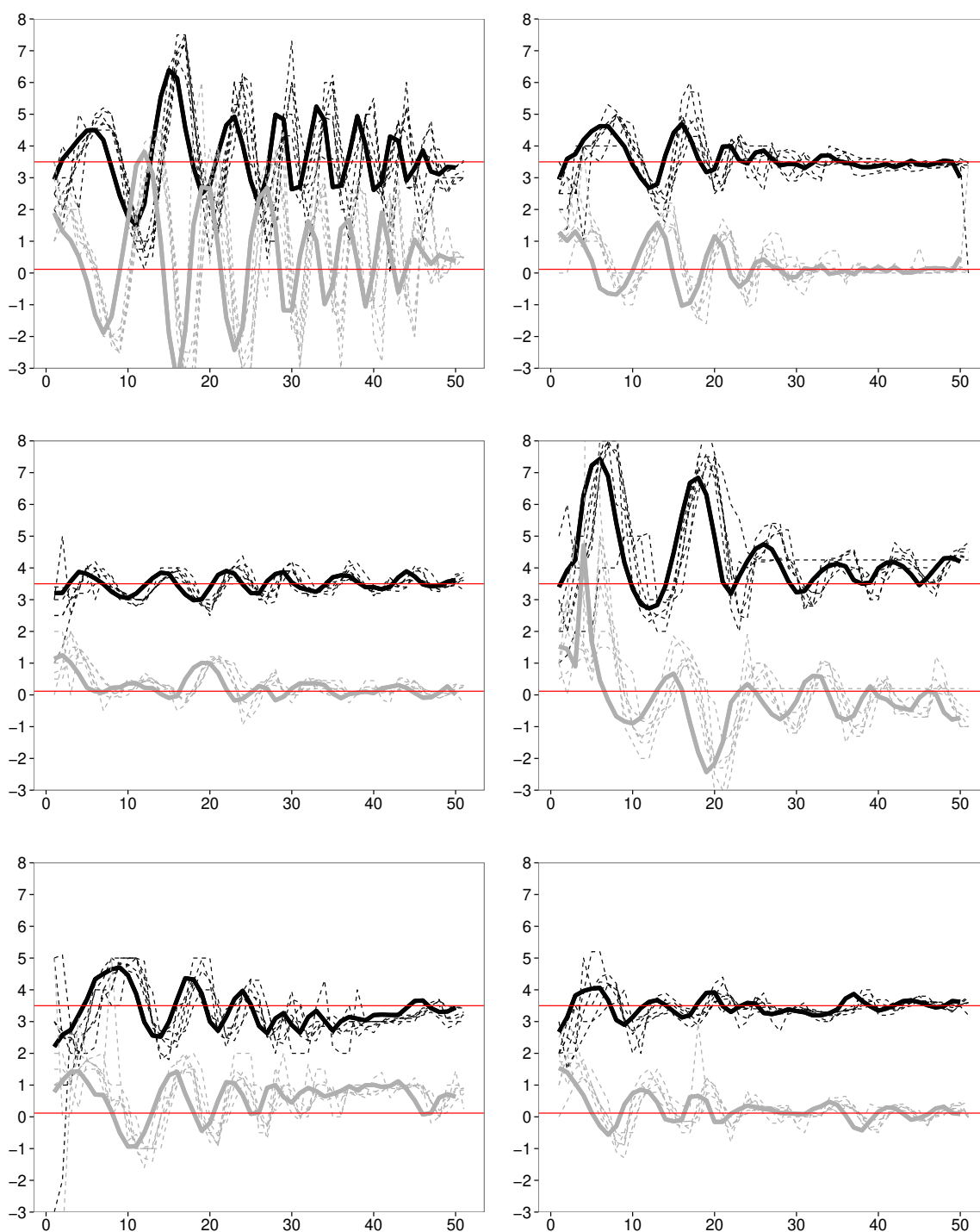


Figure 3.14: Realizations and forecasts of inflation and output gap ($T1$, groups 1 – 6)

Notes: Each of the graphs corresponds to one group and shows realized inflation (thick black line), individual inflation forecasts (dashed black lines), realized output gap (thick gray line), and individual output gap forecasts (dashed gray lines) over the 50 periods of the experiment.

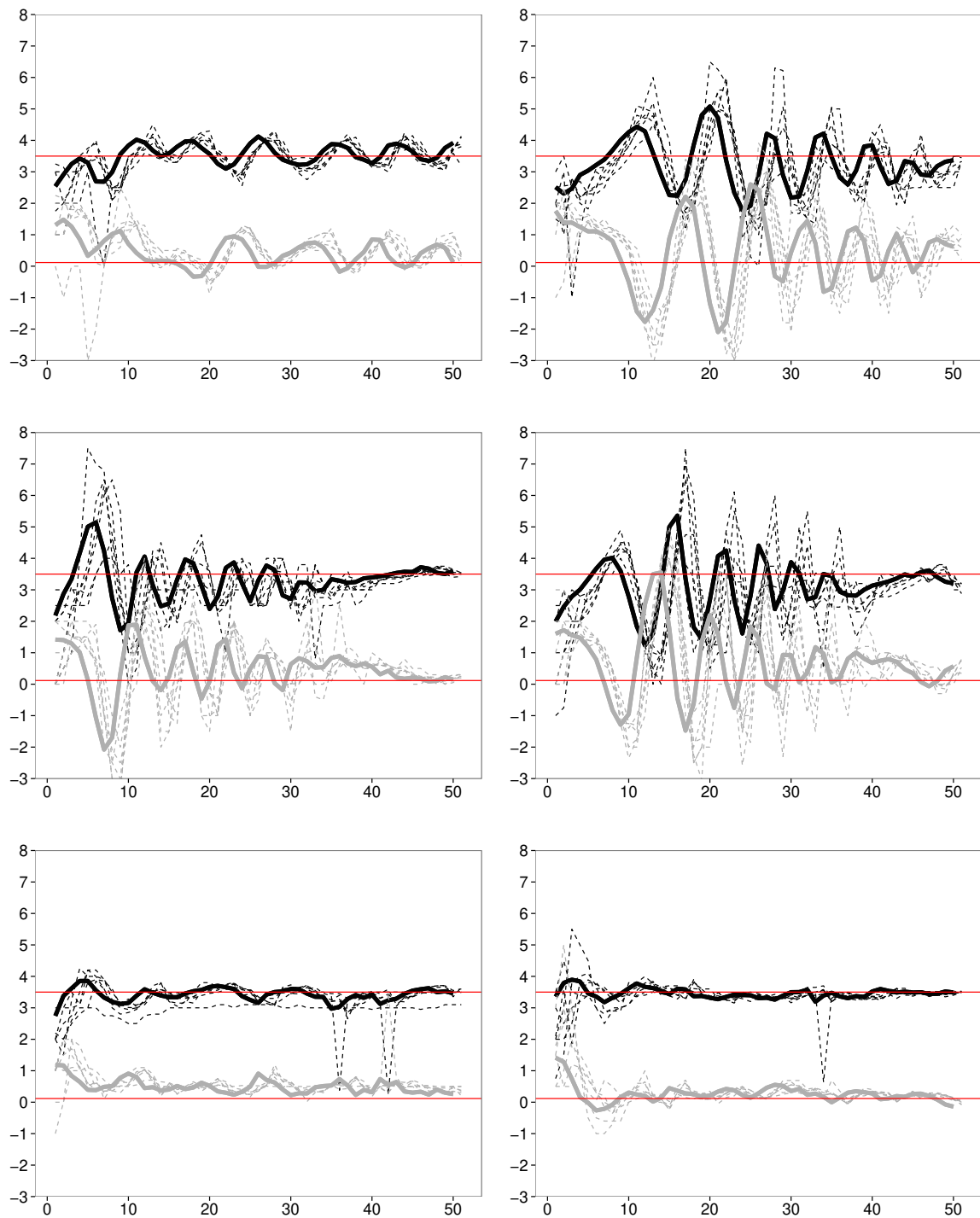


Figure 3.15: Realizations and forecasts of inflation and output gap ($T1$, groups 7 – 12)

Notes: Each of the graphs corresponds to one group and shows realized inflation (thick black line), individual inflation forecasts (dashed black lines), realized output gap (thick gray line), and individual output gap forecasts (dashed gray lines) over the 50 periods of the experiment.

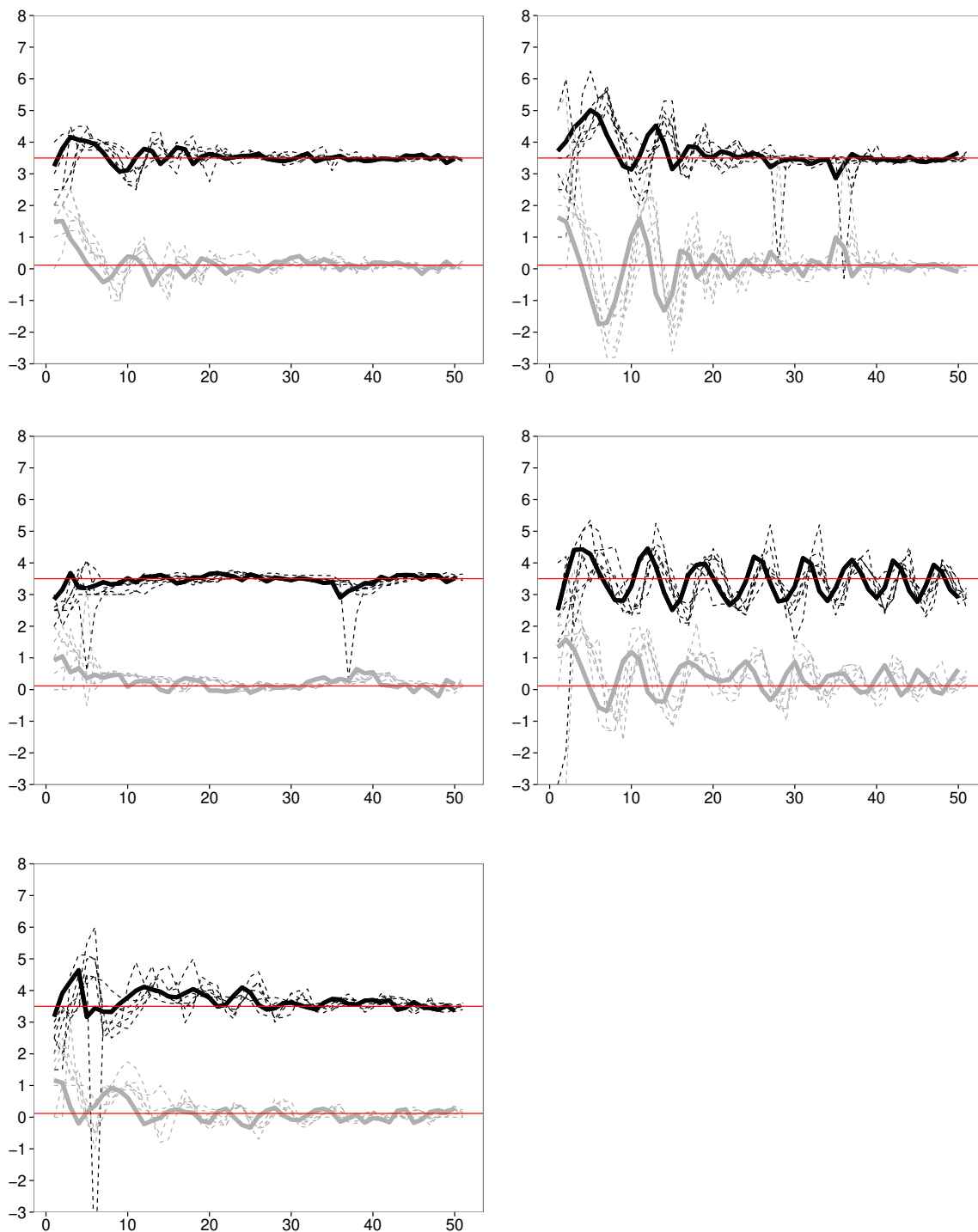


Figure 3.16: Realizations and forecasts of inflation and output gap ($T1$, groups 13 – 17)

Notes: Each of the graphs corresponds to one group and shows realized inflation (thick black line), individual inflation forecasts (dashed black lines), realized output gap (thick gray line), and individual output gap forecasts (dashed gray lines) over the 50 periods of the experiment.

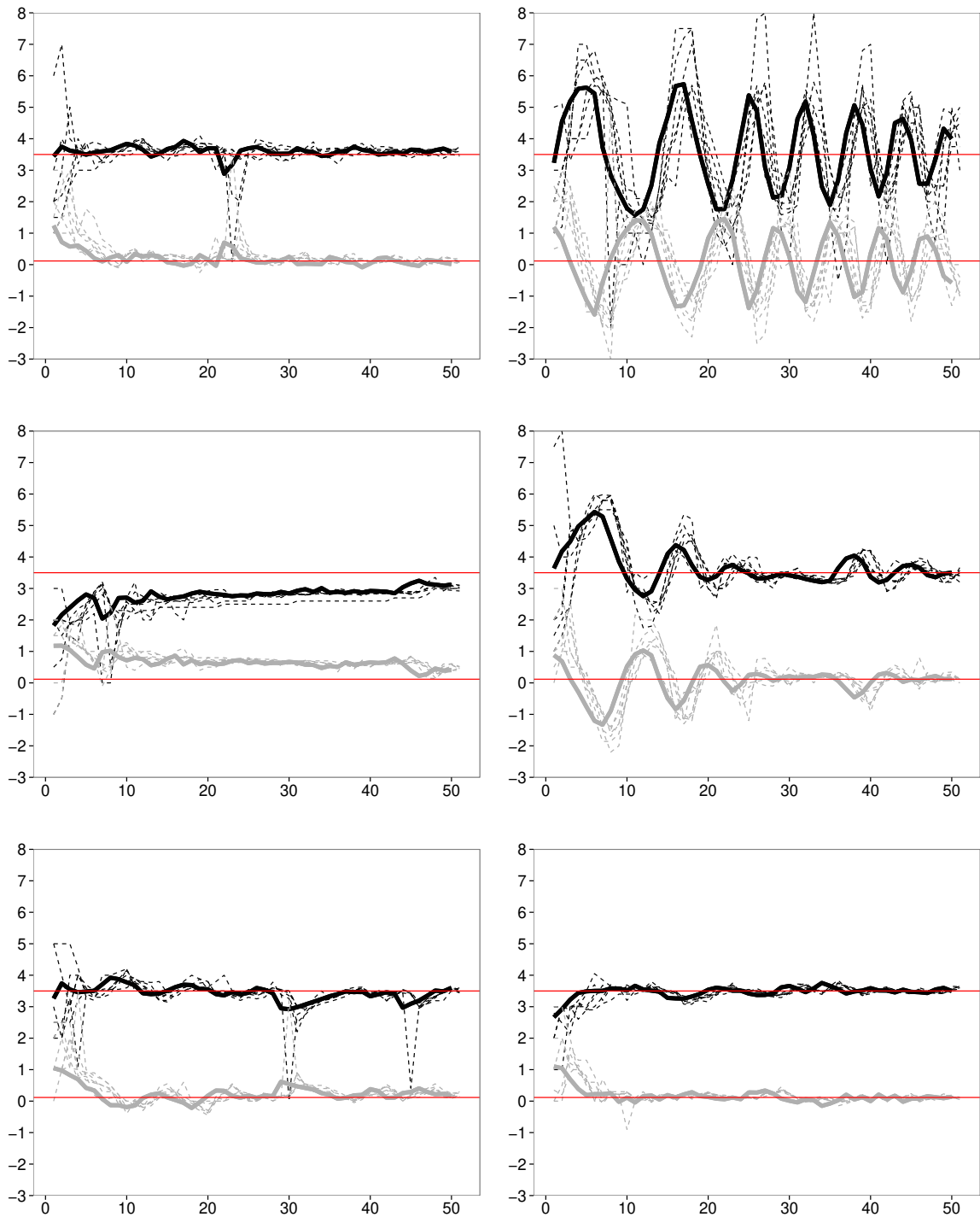


Figure 3.17: Realizations and forecasts of inflation and output gap ($T2$, groups 1 – 6)

Notes: Each of the graphs corresponds to one group and shows realized inflation (thick black line), individual inflation forecasts (dashed black lines), realized output gap (thick gray line), and individual output gap forecasts (dashed gray lines) over the 50 periods of the experiment.

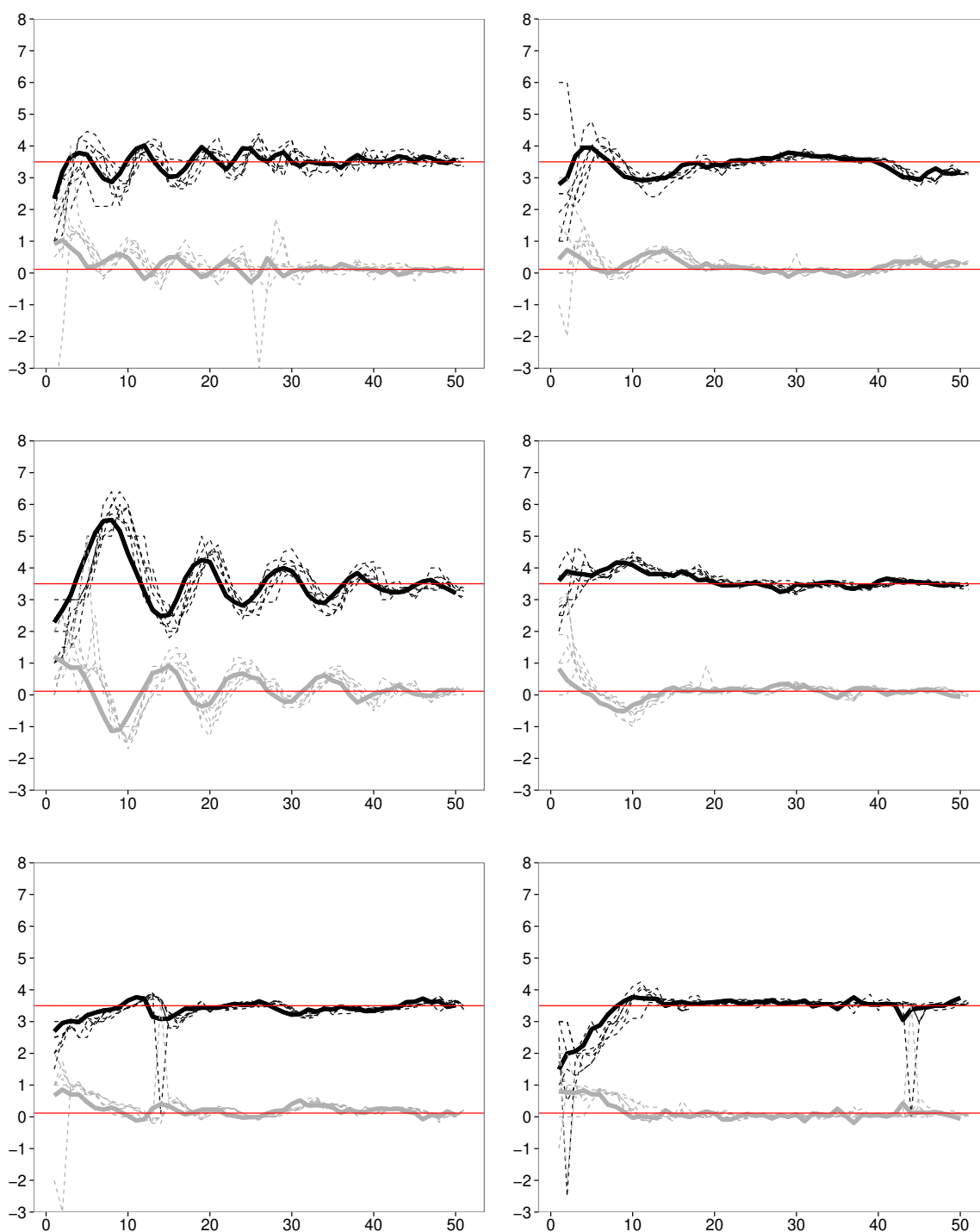


Figure 3.18: Realizations and forecasts of inflation and output gap (T^2 , groups 7 – 12)

Notes: Each of the graphs corresponds to one group and shows realized inflation (thick black line), individual inflation forecasts (dashed black lines), realized output gap (thick gray line), and individual output gap forecasts (dashed gray lines) over the 50 periods of the experiment.

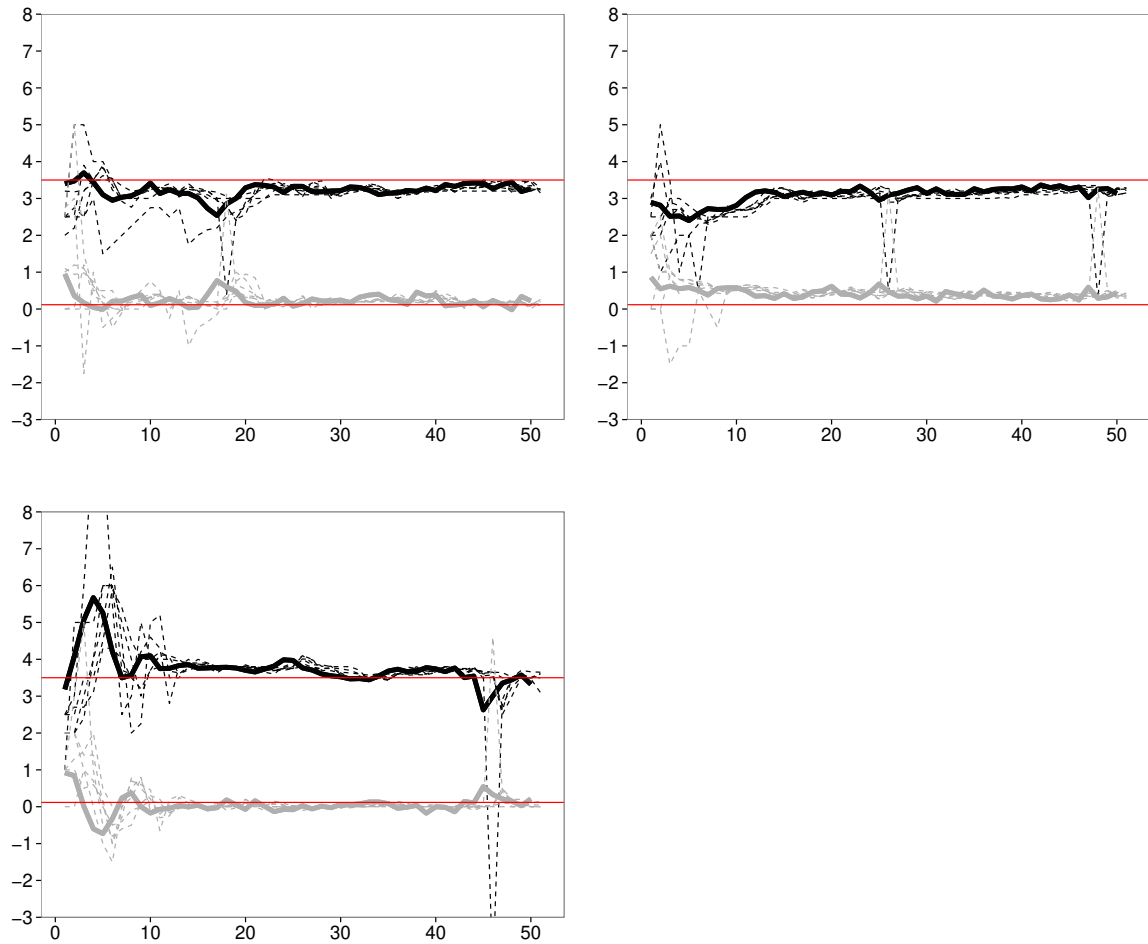


Figure 3.19: Realizations and forecasts of inflation and output gap (T^2 , groups 13 – 15)

Notes: Each of the graphs corresponds to one group and shows realized inflation (thick black line), individual inflation forecasts (dashed black lines), realized output gap (thick gray line), and individual output gap forecasts (dashed gray lines) over the 50 periods of the experiment.

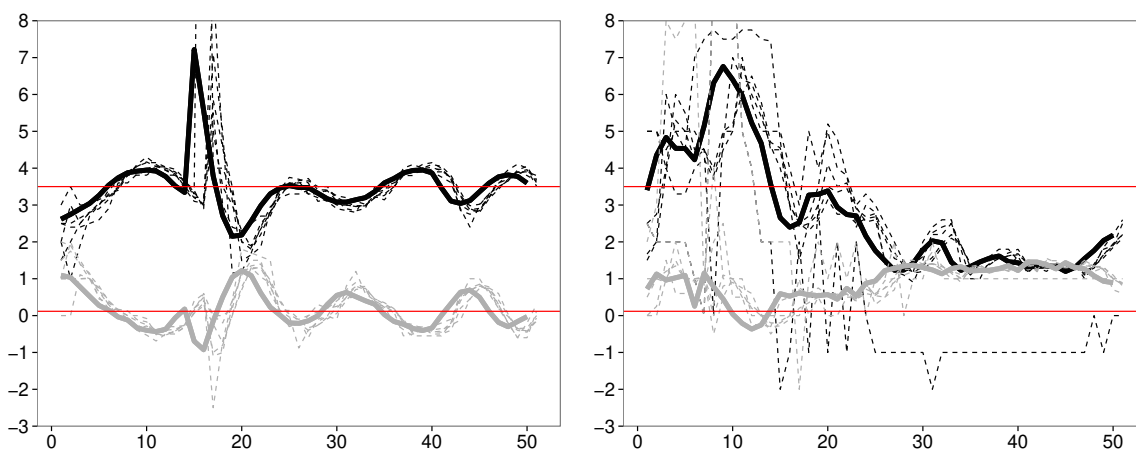


Figure 3.20: Realizations and forecasts of inflation and output gap (excluded groups)

Notes: Each of the graphs corresponds to one group and shows realized inflation (thick black line), individual inflation forecasts (dashed black lines), realized output gap (thick gray line), and individual output gap forecasts (dashed gray lines) over the 50 periods of the experiment.

3.C. Additional Graphs of the Experimental Data and Screenshot

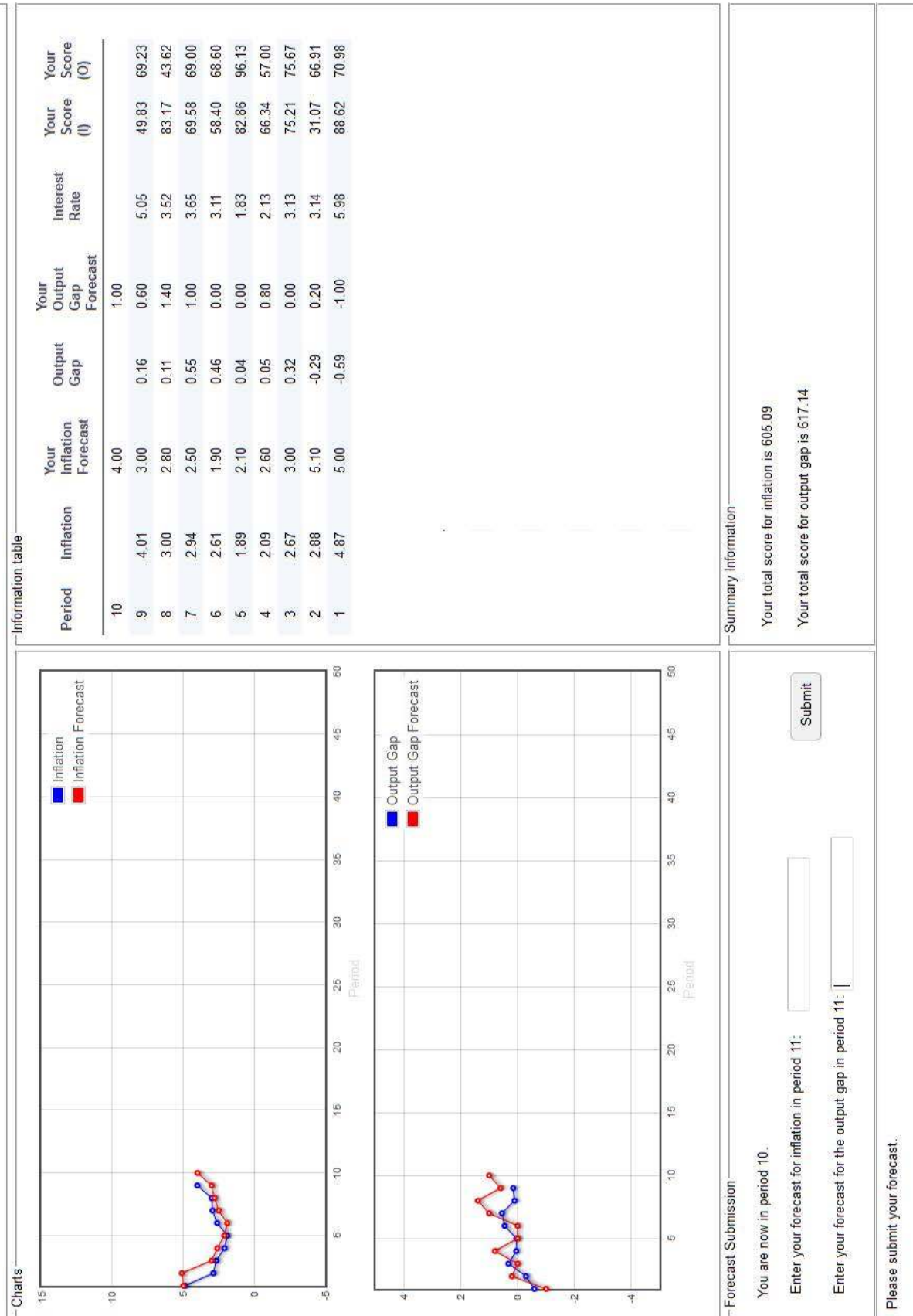


Figure 3.21: Screenshot

Chapter 4

Two-Tier Voting: Measuring Inequality and Specifying the Inverse Power Problem*

4.1 Introduction

The term two-tier voting refers to situations where different groups have to make a collective decision and do so by voting in an assembly of representatives with one representative per group. Daily, many decisions are taken through such voting by all kinds of institutions. The best-studied case is perhaps the Council of the European Union,⁴⁸ but it is by far not the only institution making use of some sort of two-tier voting. Other institutions include the UN General Assembly, WTO, OPEC, African Union, German Bundesrat, ECB, and thousands of boards of directors and professional and non-professional associations. The importance of two-tier voting is likely to further increase in the future. Globalization and the emergence of democracy in many parts of the world make collaboration in supra-national organizations more necessary and easier; furthermore, modern communication technologies facilitate the organization in interest-groups, clubs, and associations, even when the members are geographically dispersed.

The question of how such two-tier voting systems should be designed remains unsolved and can in full generality certainly not be solved. Nevertheless, there are theoretical concepts that provide guidelines, often stating which voting systems are fair. However, actual voting

*This chapter is based on Weber (2015b).

⁴⁸The literature on two-tier voting within the EU includes, among many others, Baldwin and Widgrén (2004), Beisbart et al. (2005), Felsenthal and Machover (2004), Laruelle and Valenciano (2002), Le Breton et al. (2012), Napel and Widgrén (2006), and Sutter (2000). For an overview of promising (voting) power research avenues see Kurz et al. (2015).

systems are never completely fair. It is then important to be able to measure how (un)equal a voting system is, i.e. how (un)equal the distribution of influence (or another variable of interest) is that a voting system generates. The inequality measure can then be used to compare voting systems within or across different populations. Such a measure could for example be used to investigate to what extent the inequality of voting systems correlates with other variables, such as income or crime rates. Furthermore, in some cases a voting system that is less equal than another one may have some advantages over the more equal one (for example it could be easier to explain its rules to citizens or this voting system could be more easily accepted by the people governed by it). It can then be important to be able to quantify by how much one voting system is more unequal than another one. I suggest to use the coefficient of variation to measure inequality in such voting settings. It can be applied to different variables of interest, such as indirect voting power (as measured by different power measures), the probability of a citizen's preferences to coincide with the voting outcome, or the number of representatives per citizen in an apportionment context.

Usually, no voting system exists that perfectly implements one of the abstract normative rules on the design of voting systems. The problem of finding voting systems that approximate these theoretical rules is called the inverse (power) problem. To specify the inverse problem, a measure is needed stating how well a voting system corresponds to a theoretical rule. I propose to use the coefficient of variation for this and compare it to using different objective functions.⁴⁹ I do this in a setting assuming that equal indirect Banzhaf voting power is desired. However, the coefficient of variation can also be applied in many other settings.

This chapter is organized as follows. In Section 4.2, I describe one of the possible rules prescribing the distribution of power that is desirable (Penrose's Square Root Rule), which can then be used in the remainder for illustrations. In Section 4.3, I discuss what properties an inequality measure for voting systems should satisfy and why the coefficient of variation is an appropriate choice. In Section 4.4, I describe how the inverse power problem can be specified and discuss how this can be done based on the coefficient of variation. In Section 4.5, I illustrate the differences between using different inequality measures and objective functions. Section 4.6 concludes.

⁴⁹I do not intend to develop algorithms solving the inverse power problem computationally given theoretical rule and measure of correspondence. This is what most of the literature does. Finding concrete solutions to the inverse power problem is not trivial; see for example Alon and Edelman (2010), De et al. (2012), Fatima et al. (2008), Kurz (2012), Kurz and Napel (2014), Leech (2003), and De Nijs and Wilmer (2012).

4.2 One Theoretical Concept: Penrose's Square Root Rule

In this section, I introduce one theoretical, abstract rule on how voting systems should be designed, called Penrose's Square Root Rule. I will use this rule as an example in the next sections.⁵⁰

There are N different groups, numbered from 1 to N , each group i consists of n_i individuals, numbered from 1 to n_i . Voting is binary, i.e. a proposal can either be accepted or rejected. Each individual favors the adoption of a proposal with probability one half, independently of all other individuals. Majority voting takes place within each group and the outcome determines the vote of the representative. The representatives of all groups come together in an assembly and it is determined according to their votes in combination with the voting system in the assembly of representatives whether a proposal is adopted or rejected.

Penrose's Square Root Rule. *The voting power of (the representative of) a group as measured by the Banzhaf index should be proportional to the square root of its population size.*

The main idea of this rule is to make it equally likely for each individual to influence the overall outcome of the two-tier voting procedure, independently of the group she belongs to. If a winning coalition turns into a losing coalition when voter j is excluded we say that voter j has a swing. The absolute Banzhaf index of a voter j is defined as the number of possible winning coalitions that turn into losing coalitions without voter j , divided by the total number of possible coalitions.⁵¹ The normalized or relative Banzhaf index is the absolute Banzhaf index normalized so that the sum of the indices of all voters equals one.

Denote by Ψ_i^B the absolute Banzhaf power index of an individual in group i arising from majority voting in this group and by Φ_i^B the absolute Banzhaf power index of group i in the assembly of representatives, which depends on the voting system in place. Then the probability that an individual in group i has a swing with respect to the overall outcome of the voting procedure (i.e. that she influences with her vote within the group the overall outcome) is Ψ_i^B times Φ_i^B , which is called the indirect Banzhaf voting power. Thus the probability of influencing the overall outcome is equal for all individuals if $\Psi_i^B \Phi_i^B$ is equal for all individuals or equivalently if

$$(4.1) \quad \Psi_i^B \Phi_i^B = \alpha$$

⁵⁰I use the most prominent rule on how two-tier voting systems should be designed, but using this rule as illustration does not mean that I endorse it as a normative concept. There are different possible criticisms of this rule, see for example Laruelle and Valenciano (2008). Furthermore, people do not necessarily like voting systems designed according to this rule (this will be investigated in Chapter 5).

⁵¹In the scenario described here, the absolute Banzhaf index of a voter is the probability that this voter has a swing.

for some constant $\alpha > 0$ and all i .⁵² It can easily be shown that equation (4.1) holds for all i if the normalized Banzhaf index of each group i is equal to

$$\frac{\frac{1}{\Psi_i^B}}{\sum_{j=1}^N \frac{1}{\Psi_j^B}}.$$

The normative rule on how to design voting systems as described here states that the indirect voting power $\Psi_i^B \Phi_i^B$ should be equal for all individuals independently of which group they are in, i.e. that equation (4.1) should hold for all i .⁵³

4.3 Measuring the Inequality of Voting Systems

Voting systems in assemblies of representatives are in general not completely fair. Sometimes one may want to quantify how unequal a voting system is. Thus, an inequality measure for a voting system \mathcal{W} in a population consisting of N groups with in total $m = \sum_{i=1}^N n_i$ individuals is needed. I assume that there is a variable of influence/representation at the individual level $r = (r_1, \dots, r_m)$ with all $r_i \geq 0$ and at least one r_i strictly positive. This variable could for example be indirect Banzhaf power as described in Section 4.2 so that $r = \Psi^B \Phi^B = (\Psi_1^B \Phi_1^B, \dots, \Psi_m^B \Phi_m^B)$. This variable could also be something different, such as for example indirect Shapley-Shubik power or the probability of being successful corresponding to expected payoffs (see e.g. Laruelle and Valenciano, 2008). It could also be the number of representatives per citizen in an apportionment context as for example for the US House of Representatives (see Balinski and Young, 2001). Then one can define the measure of inequality as generated by a two-tier voting system (with a very slight abuse of notation) as $\lambda(\mathcal{W}, n_1, \dots, n_N) := \lambda(r)$.

Such an inequality measure $\lambda(r)$ should satisfy certain axioms (for general treatments of inequality measures see e.g. Atkinson, 1970, or Cowell, 2011). Important axioms are:

Anonymity: $\lambda(r_1, \dots, r_m) = \lambda(r_{k_1}, \dots, r_{k_m})$ for any permutation (k_1, \dots, k_m) of $(1, \dots, m)$. This axiom states that all individuals are equally important for the inequality measure.

Scale Invariance: $\lambda(r) = \lambda(\gamma r)$ for any $\gamma > 0$. This axiom states that the unit of measurement of influence (or representation) should not matter for the inequality measure.

Population Principle: $\lambda(r_1, \dots, r_m) = \lambda(\overbrace{r_1, \dots, r_1}^k, \overbrace{r_2, \dots, r_2}^k, \dots, \overbrace{r_m, \dots, r_m}^k)$. This axiom

⁵²It is assumed that the grand coalition, i.e. all representatives voting together, can always pass a proposal. This excludes the trivial case $\alpha = 0$.

⁵³The reason why this is usually referred to as square root rule is the following. Ψ_i^B in equation (4.1) can be approximated by $\sqrt{\frac{2}{\pi n_i}}$, thus equation (4.1) holds if the Banzhaf indices of the groups are proportional to the square root of population size.

states that if a population is an identical multiplication of another one with respect to the influence each individual has, both populations (with their voting systems) should be judged to be equally unequal.

Principle of transfers: $\lambda(r_1, \dots, r_{k_i}, \dots, r_{k_j}, \dots, r_m) > \lambda(r_1, \dots, r_{k_i} + h, \dots, r_{k_j} - h, \dots, r_m)$ for any $h > 0$ and $i, j \in \{1, \dots, m\}$ with $r_{k_i} + h \leq r_{k_j} - h$. This axiom states that the inequality measure should decrease if one can decrease the influence by one citizen by a bit while simultaneously increasing the influence by another citizen who has less influence by the same amount (assuming that the redistribution does not change the ordering of influence between these two citizens).

There are multiple well-known inequality measures that satisfy these axioms (e.g. the Gini coefficient, the coefficient of variation, or the Theil index). I will focus here on the coefficient of variation and argue that it is a good choice to measure inequality in such voting settings.

The coefficient of variation is defined as the ratio of the (population) standard deviation σ to the (population) mean μ , $cv = \frac{\sigma}{\mu}$, thus in our case

$$cv(r) = \frac{\sqrt{\frac{1}{m} \sum_{i=1}^m (r_i - \bar{r})^2}}{\bar{r}}.$$

It is thus the inverse of the signal-to-noise ratio. The coefficient of variation satisfies all of the axioms stated above and is a straightforward, easily understandable inequality measure. It is a big advantage of the coefficient of variation over other measures satisfying these axioms (such as for example the Theil index) that it is so simple and easily understandable. Some researchers and certainly most policy makers or politicians concerned with the design of voting systems are not specialists in inequality measurement; therefore having such a straightforward measure is certainly good.

A further property of the coefficient of variation is that redistributing influence at any end of the distribution reduces (or increases) the inequality measure by the same amount, which can be seen as an advantage over the Gini index, another widely used measure of inequality.⁵⁴ On the opposite side, an advantage of the Gini index contributing to its popularity in measuring income or wealth inequality is the fact that it can account for negative values (such as debt); however, this advantage of the Gini index plays no role when measuring the inequality of a voting system, because influence/representation is generally non-negative. In

⁵⁴More precisely, an infinitesimal transfer from an individual with influence y_1 to an individual with influence $y_1 - h$ will always have the same effect on the coefficient of variation independently from where y_1 lies in the distribution. For the Gini coefficient, this effect depends on the distribution with usually larger effects in the middle of the distribution than in the tails, which does not constitute a particularly desired property (see Atkinson, 1970).

summary, the coefficient of variation appears to constitute a good inequality measure for voting systems.

4.4 The Inverse Power Problem

If one wants to find a voting system which optimally approximates equal indirect voting power, the inverse power problem needs to be solved. This problem is specified with an error term (or objective function) describing how much a voting system deviates from equal indirect voting power.⁵⁵ I first describe possible error terms and then discuss specifying the inverse problem with the coefficient of variation. In this section, I assume that indirect Banzhaf power is the variable of interest, but using the coefficient of variation is by no means restricted to such a setting (applying the method to other settings is straight-forward; the relation to apportionment is made in Footnote 57).

4.4.1 Error Terms Based on Voting Power on the Group Level

The system of equations (4.1) usually does not hold exactly for any voting system. It is thus necessary to find a voting system approximating full equality. One way to do this is to take a voting system that minimizes the deviation of the normalized Banzhaf index of each group from the vector that would yield equal indirect voting power. One can then take the euclidean distance as error term (i.e. as objective function) or equivalently its square, which leads to the minimization of

$$(4.2) \quad err_{group,basic}(\Psi^B, \Phi^B) := \sum_{i=1}^N \left(\frac{\Phi_i^B}{\sum_{j=1}^N \Phi_j^B} - \frac{\frac{1}{\Psi_i^B}}{\sum_{j=1}^N \frac{1}{\Psi_j^B}} \right)^2.$$

Such a squared error term at the group level has been frequently used in the literature.⁵⁶

It is easily seen that this cannot be the best term to minimize. The groups have different sizes and the idea is to equalize voting power at the individual level. I will now propose a first way to fix this (as I will show later on, the easy fix does not work perfectly). This easy fix consists of weighing the squares in the error term by their group size. In order for this error term not to increase with the number of groups or the group sizes, one can divide

⁵⁵It is of course possible to address the inverse problem only within a subset of voting systems. Such a subset could for example be all weighted voting systems, all weighted voting systems satisfying some additional conditions (e.g. the quota can be at most two thirds), or all double majority voting systems.

⁵⁶See for example Barthél my and Martin (2011), Kirsch and Langner (2011), Leech (2002), Turnovec (2011), or ˙Zyczkowski and Słomczyński (2013). Note that the main scientific contributions of these works are not corrupted by using this suboptimal error term.

by the total number of individuals. One can furthermore take the square root, such that the error term is measured in the ‘unit’ of indirect voting power rather than in its square (taking the square root only changes things if one is interested in relative differences, otherwise the square root can be left out as is usually done in equation 4.2). This leads to the minimization of

$$(4.3) \quad err_{group,imp}(\Psi^B, \Phi^B) := \sqrt{\frac{1}{\sum_{i=1}^N n_i} \sum_{i=1}^N n_i \left(\frac{\Phi_i^B}{\sum_{j=1}^N \Phi_j^B} - \frac{\frac{1}{\Psi_i^B}}{\sum_{j=1}^N \frac{1}{\Psi_j^B}} \right)^2}.$$

4.4.2 An Error Term Based on Normalized Indirect Voting Power

Another error term sometimes used in the literature (e.g. Le Breton et al., 2012, Maaser and Napel, 2007) is as follows. Rather than deriving the power distribution at the group level that leads to equal indirect voting power at the individual level, one considers indirect voting power $\Psi_i^B \Phi_i^B$ directly. One then normalizes this index of indirect voting power, so that it sums up to one when added up over all individuals. This yields a ‘normalized indirect voting power index’ of the form

$$\frac{\Psi_i^B \Phi_i^B}{\sum_{j=1}^N n_j \Psi_j^B \Phi_j^B}.$$

Then one chooses the voting system that minimizes the sum of the squared deviations of this index from one over the number of individuals, so that one ends up minimizing

$$(4.4) \quad err_{indirect}(\Psi^B, \Phi^B) := \sum_{i=1}^N n_i \left(\frac{\Psi_i^B \Phi_i^B}{\sum_{j=1}^N n_j \Psi_j^B \Phi_j^B} - \frac{1}{\sum_{j=1}^N n_j} \right)^2.$$

Again, one could take the square root (yielding the euclidean distance of the normalized indirect power vector from $1/\sum_{j=1}^N n_j$), but it is usually left out.

4.4.3 Using the Coefficient of Variation to Specify the Inverse Problem

Starting out a bit differently, the following way to specify the inverse power problem seems natural. One is looking for a voting system where indirect voting power is as equal as possible. I propose to choose the voting system that directly minimizes the inequality of indirect voting power, thus $\lambda(\Psi^B \Phi^B)$ for an inequality measure λ . Potentially any inequality measure could be used here, such as for example the Gini index. However, as argued in Section 4.3, the coefficient of variation is an appropriate measure of inequality for voting systems, thus an intuitive procedure is to minimize $cv(\Psi^B \Phi^B)$.

Now, I briefly show how using the coefficient of variation can also be derived in a similar

way to motivating the objective functions above. If the system of equations (4.1) holds, all individuals have equal (indirect) voting power. Keeping in mind that the error at the individual level is what we are interested in, one can then minimize

$$(4.5) \quad \sum_{i=1}^N \sum_{j=1}^{n_i} (\Psi_i^B \Phi_i^B - \alpha)^2 = \sum_{i=1}^N n_i (\Psi_i^B \Phi_i^B - \alpha)^2.$$

over different voting systems. As equal indirect voting power corresponds to equation (4.1) holding for *any* $\alpha > 0$ it is natural to give each voting system its ‘best shot’, i.e. to let α depend on the voting system (so that both α and Φ^B depend on the voting system):

$$\alpha = \arg \min_{\gamma} \sum_{i=1}^N n_i (\Psi_i^B \Phi_i^B - \gamma)^2.$$

It can easily be shown that then

$$(4.6) \quad \alpha = \frac{1}{\sum_{i=1}^N n_i} \sum_{i=1}^N n_i \Psi_i^B \Phi_i^B =: \overline{\Psi^B \Phi^B}.$$

Note that $\overline{\Psi^B \Phi^B}$ is the mean of $\Psi^B \Phi^B$ (taken at the individual level). Minimizing expression 4.5 with α as in 4.6 can still be adjusted. To make the error term independent of the number of groups and the group sizes, one can divide by the number of individuals. Furthermore, as for $err_{group,imp}$, if one wants to measure the variation of indirect voting power in the same unit as indirect voting power rather than its square, one can take the square root. Finally, it is desirable that the scale used does not change the relevant expressions, i.e. that merely multiplying the indices Φ^B of all groups with a constant does not change the outcome. This can be achieved by dividing through $\overline{\Psi^B \Phi^B}$. It turns out that this error term is then equal to the coefficient of variation of indirect voting power:

$$(4.7) \quad cv(\Psi^B \Phi^B) := \frac{\sqrt{\frac{1}{\sum_{i=1}^N n_i} \sum_{i=1}^N n_i (\Psi_i^B \Phi_i^B - \overline{\Psi^B \Phi^B})^2}}{\overline{\Psi^B \Phi^B}}.$$

It can be shown that for any given population (for the same N, n_1, \dots, n_N) $err_{indirect}$ is just a monotonic transformation of cv . This means that when addressing the inverse problem, either of the two leads to the same results. This is not self-evident, if one were to choose a different inequality measure this result would in general not hold. However, as I have argued, the coefficient of variation is a good inequality measure for voting systems; this equivalence can thus be seen as support for the results from using $err_{indirect}$. Note, however,

that $err_{indirect}$ is not appropriate to measure inequality across different populations.⁵⁷

4.5 Illustrations

In this section I illustrate with four (hypothetical) examples that the coefficient of variation is suitable to specify the inverse problem and to measure the inequality of voting systems.⁵⁸ The first two examples are concerned with the inverse problem, where using cv and $err_{indirect}$ lead to the same outcome. I show how using the coefficient of variation (or equivalently $err_{indirect}$) is to be preferred over using $err_{group,imp}$ (as argued above, $err_{group,basic}$ is clearly not optimal, therefore I do not consider it here). In the first example, the means of indirect Banzhaf voting power are equal under two voting systems, but the standard deviations are different. In the second example, the standard deviations are equal, but the means are different. In the third and fourth examples, I compare the inequality of two voting systems in different populations. Here, in contrast to using cv , using $err_{indirect}$ does not yield convincing results (similarly, the euclidean distance would not yield convincing results).

4.5.1 First Example: Mean-Preserving Spread

There are six groups, numbered from 1 to 6. Groups 1 and 2 have ten members each, the other groups have five members. This means that in the first stage (the election of the representatives) individuals have voting power $\Psi_{1,2}^B = 0.2460938$ and $\Psi_{3,4,5,6}^B = 0.375$, respectively. Indirect voting power would be equal across all individuals if the voting systems were such that

$$\frac{\Phi_{1,2}^B}{\sum_{i=1}^6 \Phi_i^B} = 0.2162162 \quad \text{and} \quad \frac{\Phi_{3,4,5,6}^B}{\sum_{i=1}^6 \Phi_i^B} = 0.1418919.$$

Now we compare two (hypothetical) voting systems \mathcal{W}_1 and \mathcal{W}_2 . The voting systems are

⁵⁷The relation to an apportionment setting where representatives per citizen is the variable of interest is as follows. Using Webster's method is equivalent to minimizing the error term $\sum_{i=1}^N n_i \left(a_i/n_i - h/\sum_{j=1}^N n_j \right)^2$ as proposed by Sainte-Laguë, where a_i is the number of seats for group/state i and h is the total number of seats to be apportioned (see Balinski and Young, 2001). This error term is the same as $err_{indirect}$ and minimizing it yields thus the same result as minimizing the coefficient of variation, which can be seen as support for Webster's method. However, using the error term proposed by Sainte-Laguë does not give a good measure to compare the inequality of different apportionments across different populations or different house sizes, for the same reasons $err_{indirect}$ does not constitute a good measure to compare inequality across populations for voting power, which will be illustrated in Sections 4.5.3 and 4.5.4.

⁵⁸All calculations in this section are straightforward. Details are available from the author on request.

such that the normalized Banzhaf indices are as follows:

$$\begin{aligned} \frac{\Phi_1^B(\mathcal{W}_1)}{\sum_{i=1}^6 \Phi_i^B(\mathcal{W}_1)} &= 0.2162162 + 0.05, & \frac{\Phi_2^B(\mathcal{W}_1)}{\sum_{i=1}^6 \Phi_i^B(\mathcal{W}_1)} &= 0.2162162 - 0.05, \\ \frac{\Phi_{3,4,5,6}^B(\mathcal{W}_1)}{\sum_{i=1}^6 \Phi_i^B(\mathcal{W}_1)} &= 0.1418919, & \text{and} \\ \frac{\Phi_{1,2}^B(\mathcal{W}_2)}{\sum_{i=1}^6 \Phi_i^B(\mathcal{W}_2)} &= 0.2162162, & \frac{\Phi_{3,4}^B(\mathcal{W}_2)}{\sum_{i=1}^6 \Phi_i^B(\mathcal{W}_2)} &= 0.1418919 + 0.05, \\ \frac{\Phi_{5,6}^B(\mathcal{W}_2)}{\sum_{i=1}^6 \Phi_i^B(\mathcal{W}_2)} &= 0.1418919 - 0.05. \end{aligned}$$

This means that for both voting systems there are groups that receive a bit more or less voting power than their fair shares. Under the first voting system, these deviations (in both directions) concern the large groups, under the second voting system, these deviations concern the small groups. Assume for simplicity and to have a nice illustration that normalized and absolute Banzhaf indices are equal. Now we can calculate the indirect voting power of each individual, depending on the group she is in. This yields

$$\begin{aligned} \Psi_1^B \Phi_1^B(\mathcal{W}_1) &= 0.06551414, & \Psi_2^B \Phi_2^B(\mathcal{W}_1) &= 0.04090477, \\ \Psi_{3,4,5,6}^B \Phi_{3,4,5,6}^B(\mathcal{W}_1) &= 0.05320946, & \text{and} \\ \Psi_{1,2}^B \Phi_{1,2}^B(\mathcal{W}_2) &= 0.05320946, & \Psi_{3,4}^B \Phi_{3,4}^B(\mathcal{W}_2) &= 0.07195946, \\ \Psi_{5,6}^B \Phi_{5,6}^B(\mathcal{W}_2) &= 0.03445946. \end{aligned}$$

One can easily see that using $err_{group,imp}$ does not distinguish between the two voting systems, both would be judged to be ‘equally equal’ ($err_{group,imp}$ equals $\sqrt{1/120}$ for both voting systems). If one looks carefully at the indirect voting power, this does not seem justified, though. For both voting systems, there are twenty individuals with indirect voting power 0.05320946, which is also the mean of indirect voting power under both voting systems. For both voting systems, there are ten individuals with higher voting power and 10 with lower power. The absolute difference between the higher value and the middle value is always equal to the difference between the middle value and the lower value; however, these differences are higher under the second voting system than under the first. The first voting system is thus less unequal than the second one. It is also selected correctly by the coefficient of variation, $cv(\Psi^B \Phi^B(\mathcal{W}_1)) = 0.1635184$ and $cv(\Psi^B \Phi^B(\mathcal{W}_2)) = 0.249171$.

4.5.2 Second Example: Shift of the Distribution

There are four groups. The first has nine members while the others have three members each. The Banzhaf power indices in the first stage are $\Psi_1^B = 0.2734375$ and $\Psi_{2,3,4}^B = 0.5$, respectively. Indirect voting power is equal for all individuals if the normalized Banzhaf indices in the assembly of representatives are

$$\frac{\Phi_1^B}{\sum_{i=1}^4 \Phi_i^B} = 0.3786982 \quad \text{and} \quad \frac{\Phi_{2,3,4}^B}{\sum_{i=1}^4 \Phi_i^B} = 0.2071006.$$

Assume again for simplicity that absolute and normalized Banzhaf indices are equal and assume that two voting systems, \mathcal{W}_1 and \mathcal{W}_2 , are such that

$$\begin{aligned} \frac{\Phi_1^B(\mathcal{W}_1)}{\sum_{i=1}^4 \Phi_i^B(\mathcal{W}_1)} &= 0.3786982 - 0.09, & \frac{\Phi_{2,3,4}^B(\mathcal{W}_1)}{\sum_{i=1}^4 \Phi_i^B(\mathcal{W}_1)} &= 0.2071006 + 0.03, \\ \text{and} & & & \\ \frac{\Phi_1^B(\mathcal{W}_2)}{\sum_{i=1}^4 \Phi_i^B(\mathcal{W}_2)} &= 0.3786982 + 0.09, & \frac{\Phi_{2,3,4}^B(\mathcal{W}_2)}{\sum_{i=1}^4 \Phi_i^B(\mathcal{W}_2)} &= 0.2071006 - 0.03. \end{aligned}$$

Which of these two voting systems is more equal? In terms of $err_{group,imp}$ the two systems approximate equal indirect voting power equally well. This can be seen as follows. The terms in parentheses in expression (4.3) are for both voting systems always either -0.09 or $+0.09$ for the parts referring to the large group and either $+0.03$ or -0.03 for the parts referring to the small groups. As only the squares of these values matter, these two voting systems are ‘equally equal’ when judged this way. Using the coefficient of variation, in contrast, makes a difference between these two voting systems. Individuals’ indirect voting power is under the first voting system

$$\Psi_1^B \Phi_1^B(\mathcal{W}_1) = 0.07894092 \quad \text{and} \quad \Psi_{2,3,4}^B \Phi_{2,3,4}^B(\mathcal{W}_1) = 0.1185503,$$

and under the second voting system

$$\Psi_1^B \Phi_1^B(\mathcal{W}_2) = 0.1281597 \quad \text{and} \quad \Psi_{2,3,4}^B \Phi_{2,3,4}^B(\mathcal{W}_2) = 0.0885503.$$

Remember that exactly half of the individuals are in the large group. Thus, under the first voting system an individual in the half of the population with more indirect voting power holds 1.50176 times as much power as an individual in the other half of the population. This ratio is only 1.447309 under the second voting system.⁵⁹ The coefficient of variation is

⁵⁹The absolute values of the differences $\Psi_1^B \Phi_1^B(\mathcal{W}_1) - \Psi_{2,3,4}^B \Phi_{2,3,4}^B(\mathcal{W}_1)$ and $\Psi_1^B \Phi_1^B(\mathcal{W}_2) -$

0.2005627 for \mathcal{W}_1 and 0.182776 for \mathcal{W}_2 . Thus, the coefficient of variation selects the second voting system, which is the one with smaller (relative) differences in indirect voting power.

4.5.3 Third Example: Comparing the Inequality of Voting Systems across Different Populations I

The first population consists of six groups of four people each. The second population consists of four groups of eight people each. Assume that the voting system in place in the first population, \mathcal{W}_X , is such that indirect voting power is

$$\Psi_{1,2,3}^B \Phi_{1,2,3}^B(\mathcal{W}_X) = 0.03 \quad \text{and} \quad \Psi_{4,5,6}^B \Phi_{4,5,6}^B(\mathcal{W}_X) = 0.01.$$

In the second population, the voting system, \mathcal{W}_Y , is such that indirect voting power is

$$\Psi_{1,2}^B \Phi_{1,2}^B(\mathcal{W}_Y) = 0.03 \quad \text{and} \quad \Psi_{3,4}^B \Phi_{3,4}^B(\mathcal{W}_Y) = 0.01.$$

This means that in the first population half of the individuals have indirect voting power of 0.03, while the other individuals have voting power 0.01. The same holds for the second population. Which of the two voting systems is more unequal? The only reasonable answer seems to be that they are ‘equally unequal’ (as stated in the population principle in Section 4.3). Using the coefficient of variation as inequality measure also yields this result, $cv(\Psi^B, \Phi^B(\mathcal{W}_X)) = 0.5 = cv(\Psi^B, \Phi^B(\mathcal{W}_Y))$. However, if one tried to use $err_{indirect}$ as a measure of inequality, one would obtain $err_{indirect}(\Psi^B, \Phi^B(\mathcal{W}_X)) = 0.0069444$ and $err_{indirect}(\Psi^B, \Phi^B(\mathcal{W}_Y)) = 0.0078125$. One would then misleadingly conclude that the voting system in the first population is more equal than the one in the second population.

4.5.4 Fourth Example: Comparing the Inequality of Voting Systems across Different Populations II

After already illustrating in the last hypothetical example that the coefficient of variation is well-suited to compare the inequality of voting systems across different populations (while for example $err_{indirect}$ is not), I will make a similar point now with an example where also the voting systems in the assemblies are specified.

In the first population there are two groups of five and two groups of three individuals. In the assembly of representatives majority voting prevails (i.e. any three representatives can pass a proposal). In the second population there are three groups of five and three groups

$\Psi_{2,3,4}^B \Phi_{2,3,4}^B(\mathcal{W}_2)$ are equal.

of three individuals. Also here, decisions in the assembly of representatives are taken by majority voting (requiring four of six representatives to pass a proposal).

This means that in the first population the voting power at the group level is $\Psi_{1,2}^B = 0.375$ and $\Psi_{3,4}^B = 0.5$. In the second population, these values are $\Psi_{1,2,3}^B = 0.375$ and $\Psi_{4,5,6}^B = 0.5$. Majority voting in the assembly of representatives leads to $\Phi_{1,2,3,4}^B(\mathcal{W}_{maj-4}) = 0.375$ in the assembly of the first population and to $\Phi_{1,2,3,4,5,6}^B(\mathcal{W}_{maj-6}) = 0.3125$ in the assembly of the second population. Thus, we have

$$\Psi_{1,2}^B \Phi_{1,2}^B(\mathcal{W}_{maj-4}) = 0.140625 \quad \text{and} \quad \Psi_{3,4}^B \Phi_{3,4}^B(\mathcal{W}_{maj-4}) = 0.1875$$

in the first population. In the second population, we have

$$\Psi_{1,2,3}^B \Phi_{1,2,3}^B(\mathcal{W}_{maj-6}) = 0.1171875 \quad \text{and} \quad \Psi_{4,5,6}^B \Phi_{4,5,6}^B(\mathcal{W}_{maj-6}) = 0.15625.$$

This means that in both populations, five eighth of the population have lower indirect voting power than the remaining three eighth. In both populations, the voting power of the part of the population with higher power is exactly a third higher than the voting power of the rest. Thus, these two population exhibit the same degree of inequality and accordingly the coefficient of variation is equal for both populations with these voting systems (it is equal to 0.1434438 in both cases). If one were to judge the inequality by $err_{indirect}$, one would conclude that the second population with its voting system is more equal (the values are 0.0012860 and 0.0008573, respectively).

4.6 Concluding Remarks

In this chapter I have first addressed the question of how the inequality generated by voting systems should be measured. I have argued that the coefficient of variation is an appropriate measure. I have furthermore argued that it is appropriate to specify the inverse power problem with the coefficient of variation when a fair voting system is desired. This is to be preferred over minimizing error terms that are based on weighted or unweighted voting power at the group level. It turns out that specifying the inverse problem with the coefficient of variation is equivalent to using an objective function based on the distance of the normalized indirect voting power from the fair ideal. Unlike the coefficient of variation, such an objective function cannot be used, however, to compare the inequality of voting systems across different populations. I have used a setting where equal indirect Banzhaf voting power is desired as illustration, but the coefficient of variation can be applied in many different settings.

The design of voting systems is a topic where policy makers and politicians are involved. It is therefore important to have salient and easily understandable tools at hand. In addition to having desired properties, the coefficient of variation is a simple and widespread statistical measure. This is an advantage over other inequality measures. Using the coefficient of variation to specify the inverse problem also seems to be more intuitive and salient for policy makers and politicians. There is no need to talk about normalizations and objective functions, one just looks straightforwardly for a voting system minimizing the inequality of the influence a citizen has on the overall outcome of a voting procedure (or any other variable of representation one is interested in).

Chapter 5

Choosing the Rules: Preferences over Voting Systems in Assemblies of Representatives*

5.1 Introduction

As already mentioned in the previous chapter, there is a variety of situations in which different groups have to make a collective decision and such a decision is often taken by voting in an assembly consisting of one representative per group. Although this topic has been widely studied, there is no agreement among scholars on the question of the design of voting systems in such an assembly.⁶⁰ There is, however, a vast literature on theoretical rules for the design of such voting systems and on the extent to which actual voting systems follow the different theoretical guidelines. Perhaps somewhat surprisingly, to date there has been no work investigating which voting systems for an assembly of representatives people actually prefer. This question is relevant because preferences for voting systems and their acceptance are closely related. It is important for people to accept the voting systems that govern them. This applies not only to the EU, where politicians are constantly concerned with the acceptance of EU institutions, but to all voting institutions, no matter whether the voting takes place at a multinational level or at a very small scale in boards or associations – given the

*This chapter is based on Weber (2015a).

⁶⁰Arguments on this topic peaked during the reformation of the EU voting system. Nine scholars wrote an open letter to the governments of the EU member states, cosigned by 38 other scholars, calling on the EU to implement Penrose's Square Root Rule (Penrose, 1946; Banzhaf III, 1964), which is the most prominent rule among scientists and policy makers (see <http://www.esi2.us.es/~mbilbao/pdf/letter.pdf>). Some government officials had already pushed for such a rule, such as the Swedish government in 2000 and, most famously, the Polish government in 2007. Despite considerable support, this rule faced opposition by other leading scholars (see, e.g., Laruelle and Valenciano, 2008, Turnovec, 2009).

large number of small organizations making use of such voting, it can be argued that the acceptance of voting systems in such organizations, associations, and boards is of particular importance. Acceptance of voting procedures is not only important for the democratic legitimacy of organizations, it can also influence outcomes; it has in general been shown that people act differently when they consider procedures fair as opposed to when they do not (see, e.g., Bolton et al., 2005, De Cremer and Tyler, 2007, Krawczyk and Le Lec, 2010, and the references therein). For all these reasons, it is surprising that the matter of preferences for such voting systems have not yet been studied. I address this gap in the literature and provide the first research investigating which voting systems for an assembly of representatives people prefer.⁶¹

A simple example serves to illustrate that it is not clear which voting system should be used in an assembly of representatives. Assume that the South American countries create an institution in which each country is represented by one representative and these representatives come together to vote on certain issues. Some people might think each country should just have one vote and the majority of votes should decide. In this case, Suriname would have the same power as Brazil, although the population of Brazil is about 400 times as large as that of Suriname. Somebody may therefore come up with the solution of weighting the votes by the population sizes and letting the majority of weighted votes decide. However, in this case, Brazil's representative could decide everything alone, because more than half of the population of South America lives in Brazil. Thus, this also does not appear to be a fair solution and the right solution seems to 'lie somewhere in between' – between giving all groups the same power and giving too much power to the largest group(s).

As mentioned in the previous chapter, prominent social choice rules concerning how voting systems in assemblies of representatives should be designed are imprecise about how the voting procedures should be carried out (e.g., 'let the majority of votes decide'). They are more abstract and prescribe how voting power should be distributed between the representatives. The two most prominent rules require either (i) that the Banzhaf power index of a representative be proportional to the square root of her group's size (Penrose's Square Root Rule) or (ii) that the Shapley-Shubik power index of a representative be proportional to her group's size. The main difference between these rules is that the first rule allocates relatively

⁶¹This research is also the first to link voting power to individuals' preferences over voting systems and furthermore the first experiment on two-tier voting. There are a few other experiments on voting power not concerned with two-tier voting, namely Montero et al. (2008), Drouvelis et al. (2010), Aleskerov et al. (2009), Guerri et al. (2014), Esposito et al. (2012), and Geller et al. (2012). Their research is carried out in bargaining settings that differ considerably from take-it-or-leave-it settings as used in this chapter. Most experimental research on voting power addresses the question how voting systems map to voting power. There is also related literature eliciting peoples' preferences for social choice rules that is not concerned with two-tier voting or voting power such as Sertel and Giritligil (2003) and Giritligil and Sertel (2005). Blais et al. (2014) investigate preferences over electoral rules; also their work is neither concerned with two-tier voting nor voting power.

more power to smaller groups than the second rule. Do people choose voting systems designed according to these concepts when they are behind the veil of ignorance, that is when they do not know which group they will be in? If so, which of these rules do they prefer? I investigate these questions in a laboratory experiment with monetary incentives. In one treatment, efficiency concerns are absent and the assumptions of the theoretical derivations of the rules are closely mimicked. In another treatment, participants' preferences over the voting outcome are perfectly correlated within a group. This allows to observe the robustness of participants' choices to variations in the assumptions underlying the rules.⁶² Furthermore, in control treatments I examine whether choices differ in front of the veil of ignorance, that is when subjects know which group they will be in.

To investigate which type of voting system people choose, a laboratory experiment is the optimal choice. In the laboratory it is possible to actually put people behind the veil of ignorance. It is also possible to imitate the assumptions made in the derivations of the normative rules (assumptions made in one treatment can then be modified in a controlled way, keeping everything else equal). Furthermore, laboratory experiments offer the possibility of incentivizing choices with monetary payments and to calculate which choices a payoff- or utility maximizing participant would make, which is in general impossible for survey data or real-world observations.

The main result from my research is the following. In the setting best mimicking the theoretical foundations of these rules (giving Penrose's Square Root Rule its best shot), participants behind the veil of ignorance prefer voting systems with proportional Shapley-Shubik power over voting systems designed according to Penrose's Square Root Rule. Voting systems designed according to Penrose's Square Root Rule are not chosen significantly more often than other voting systems that do not follow any reasonable rule. This means that participants prefer voting systems that give more voting power to larger groups than the most prominent theoretical concept prescribes. This result cannot be explained by choices according to outcome based preferences (such as risk-aversion or other-regarding preference models). Further results are that when the incentive structure is altered participants choose voting systems more often that give higher expected payoffs and that in front of the veil of ignorance participants choose voting systems that are good for their own group.

This chapter is organized as follows. In Section 5.2, I give a brief description of the rule prescribing proportional Shapley-Shubik power (Penrose's Square Root Rule has already

⁶²The normative rules are derived from particular examples but are applied very generally for real world voting institutions. It is thus interesting to see not only if these rules are accepted when their assumptions are perfectly fulfilled, but also when their assumptions are violated (theoretical research shows that the rules are sensitive to changes in the assumptions, see Kaniowski, 2008, Kurz et al., 2013b, Le Breton and Van Der Straeten, 2014; to some extent, one can see from this research to what extent this also holds for individuals' choices).

been described in Chapter 4) for designing voting systems in assemblies of representatives. Furthermore, I derive the behavior of individuals maximizing expected utility (defined over outcomes). Section 5.3 contains the experimental design and Section 5.4 contains the data analysis and the results. Section 5.5 concludes the chapter.

5.2 Equalizing Voting Power and Maximizing Utility

Penrose's Square Root Rule has already been introduced in Chapter 4. Here, I briefly present one motivation for the rule prescribing proportional Shapley-Shubik power. The outcomes of the two rules can be quite different, the main difference being that Penrose's Square Root Rule allocates relatively more power to smaller groups than the rule prescribing proportional Shapley-Shubik power (this in general holds no matter how power is measured). After presenting the second rule, I show how utility maximizing agents choose voting systems in a take-it-or-leave-it two-tier voting setting (where utility is defined over outcomes).⁶³

Again, the underlying situation is the following. There are N different groups, numbered from 1 to N . Each group i consists of n_i individuals, numbered from 1 to n_i . Each group elects one representative through majority voting (in the first or 'lower' tier). The representatives then come together in an assembly to vote (in the second or 'upper' tier). The representatives vote on an issue concerning all individuals in the best interest of their group. The voting system governing the voting in this assembly of representatives is the focus of most of the two-tier voting literature.⁶⁴ Note that there are other voting systems than those that can be represented by weighted voting; for example double majority systems, as used in the Council of the EU, can generally not be implemented by just giving the votes different weights.

The two rules under consideration were developed as normative rules. This research investigates the choices people make; thus, it examines two-tier voting from a positive per-

⁶³The vast majority of the two-tier voting literature is not based on utility but on voting power. Although there may be agreement between utility and voting power considerations (particularly voting in bargaining committees), they are generally different. In take-it-or-leave-it settings, utility-based concepts have emerged only very recently (Barbera and Jackson, 2006; Beisbart et al., 2005; Koriyama et al., 2013; Laruelle and Valenciano, 2010). One reason why the focus was previously mainly on voting power may be that it can be derived from a voting system alone without specifying utility. Both, when voting power and utility considerations do and when they do not coincide, there are opportunities to conduct further meaningful research involving voting power (Kurz et al., 2015). As the scientific community has focused on voting power rather than utility for such a long time (as a normative concept) it seems very natural that voting power plays an important role in preferences over voting systems.

⁶⁴Often, the terms 'voting system' and 'voting rule' are used interchangeably. Here, I use 'voting rule' for an abstract rule describing voting systems, which can be applied in different situations, i.e., different numbers of groups and individuals per group (such as Penrose's Square Root Rule). I use the term 'voting system' when the number of groups and the number of individuals per group are fixed.

spective. Proponents of these rules may argue that they never claimed that people would choose according to these concepts. Nevertheless, there is a clear connection between the normative side and the positive side. If people are put behind the veil of ignorance, do they choose voting systems recommended by these rules? If so, which of these concepts do they favor? How robust are their choices to violations of the assumptions underlying the theories?

The theoretical rules cannot be implemented perfectly. Therefore, the inverse power problem needs to be resolved (as discussed in Chapter 4). For all constellations of voting systems that I use in the experiment the different methods based on $err_{group,imp}$ and cv (cf. equations 4.3 and 4.7) yield the same unique outcome. If one of the two methods is rejected on a theoretical basis, this has thus no consequences for the conclusions of this experiment.

5.2.1 Penrose's Square Root Rule and Proportional Shapley-Shubik Power

As a reminder, Penrose's Square Root Rule, the most prominent normative concept of how two-tier voting systems should be designed, is as follows (see Section 4.2).⁶⁵

Rule I (Penrose's Square Root Rule). *The voting power of (the representative of) a group as measured by the Banzhaf index should be proportional to the square root of its population size.*

Another prominent normative concept for the design of two-tier voting systems is as follows. Much of the derivation for the probabilistic motivation of this rule can be performed similarly to Section 4.2. The motivation is thus kept brief.⁶⁶

Rule II (Proportional Shapley-Shubik Power). *The voting power of (the representative of) a group as measured by the Shapley-Shubik index should be proportional to its population size.*

In contrast to the derivation of Penrose's Square Root Rule, it is now assumed that all voters differ in the strength of their feelings over the issue at stake. One can then order all voters from strong like to strong dislike. In general, voter j is in a pivotal position if the coalition of voters that would like the adoption of a proposal more strongly than voter j does not have the power to pass it, whereas the coalition of voters that would like the adoption of the proposal less (dislike it more) does not have the power to block it. A voter in a pivotal position is thought to have decisive influence over the outcome of the voting process.

⁶⁵I will not use the square root approximation (see Footnote 53) in this chapter, but still talk about the Banzhaf index being proportional to the square root of group size (working with the exact value or the approximation makes no conceptual difference).

⁶⁶The probabilistic motivation (for a slightly more detailed description see Turnovec, 2009, and Turnovec et al., 2008) is not the only motivation for this rule; it can also be motivated in a bargaining committee setting. The Shapley-Shubik index originates from cooperative game theory (Shapley and Shubik, 1954, Shapley, 1953).

I now state the relevant definitions, in accordance with the literature. Let (i_1, \dots, i_M) be a permutation of voters (voters are numbered from 1 to M ; the voting system – i.e., the set of winning coalitions – is denoted by \mathcal{W}). If voter j 's position in the permutation is i_k , then voter j is pivotal if $\{i_1, \dots, i_{k-1}\} \notin \mathcal{W}$ and $\{i_1, \dots, i_k\} \in \mathcal{W}$. The Shapley-Shubik power index of voter j is the number of permutations in which j is pivotal divided by the total number of permutations $M!$. Note that the sum of the Shapely-Shubik indices of all voters equals one and that this index represents the probability of being pivotal if all permutations (that can be seen as preference orderings) are equally likely.

Denote by Ψ_i^S the Shapley-Shubik power index of an individual in group i arising from majority voting and by Φ_i^S the Shapley-Shubik index of group i in the assembly of representatives, depending on the voting system in the assembly. Assuming that all permutations are equally likely in both stages of the voting procedure, the probability that an individual in group i is pivotal in the first stage while the representative of group i is pivotal in the second stage is $\Psi_i^S \Phi_i^S$. Then, the probability of influencing the overall outcome is equal for all individuals if

$$(5.1) \quad \Psi_i^S \Phi_i^S = \alpha$$

for all i and some constant α . Because the Shapley-Shubik indices of all voters sum to one, it is $\Psi_i^S = \frac{1}{n_i}$. Thus, equation (5.1) holds for all i if the Shapley-Shubik index of each group i is equal to $\frac{n_i}{\sum_{j=1}^N n_j}$, i.e., if the Shapley-Shubik indices of the groups are proportional to their sizes.

5.2.2 Choosing Voting Systems According to Expected Utility Theory

Now, instead of considering recommendations for how voting systems should be designed, I derive the way that utility maximizing individuals would choose voting systems in a setting that can be applied one-to-one to the experiment. To be able to derive predictions for the choice of voting systems by utility maximizing, it is necessary to specify payoffs. The most natural specification (in particular in view of the derivation of Penrose's Square Root Rule) is a fixed payment to an individual if the overall voting outcome coincides with her preferences with respect to the outcome.⁶⁷ Without loss of generality, one can assume that any individual has utility function u , normalized to 1 if the outcome coincides with her preferences and to 0 if it does not.

Assume that only one issue is voted on. Because there is majority voting at the group

⁶⁷In the experiment, individuals who are in favor of adopting a proposal obtain a payment of 1000 points (12.50 euros) if the proposal is adopted and a payment of 0 otherwise. Individuals in favor of rejecting a proposal obtain a payment of 1000 points if the proposal is rejected and 0 otherwise.

level and the representative acts in the best interest of her group, individuals can only influence their own (expected) payment through the choice of the voting system.⁶⁸

In general, different interpretations of the veil of ignorance are possible. The interpretation that is appropriate for this experiment is unambiguous, however. Being behind the veil of ignorance means not knowing which group one will be in and, more precisely, not knowing which individual one will be. Each individual is in the i -th group g_i with probability proportional to its size, $p(g_i) = \frac{n_i}{\sum_{i=1}^N n_i}$. These probabilities are known.

Let \mathbf{W} denote the set of voting systems (where a voting system is fully characterized by an exhaustive list of winning coalitions) from which the individual can choose and u her utility after the voting procedure has been applied (once). A utility maximizing agent chooses a voting system as follows (being ‘successful’ means that the preferred outcome coincides with the actual voting outcome):

$$(5.2) \quad \begin{aligned} \mathcal{V}_{B,max} &= \arg \max_{\mathcal{W} \in \mathbf{W}} E(u|\mathcal{W}) \\ &= \arg \max_{\mathcal{W} \in \mathbf{W}} \sum_{i=1}^N p(g_i) p(\text{‘individual of } g_i \text{ successful’}|\mathcal{W}). \end{aligned}$$

This means that a utility maximizing individual in the experiment chooses the most efficient voting system. The probability of success for an individual in a certain group depends on the voting system as well as on how preferences are formed. The assumption of independently drawn voting outcome preferences for each individual (where everyone is equally likely to favor the adoption or rejection of a proposal independently of everyone else) can be used as well as any other specification of probabilities or correlation structures. For any given voting system and assumptions governing the probability distribution, the expected utility can be calculated (or simulated).

This choice would be made by a utility maximizing economic agent and – in case the voting procedure is only performed once (i.e., only one issue is voted on) – by anyone with standard outcome-based preferences (exhibiting, e.g., altruism or aversion to inequality). At the same time, this choice would be made by a social welfare maximizer with a utilitarian social welfare function (and with many other social welfare functions). This can be seen as follows. Only one issue is voted on, and only one voting system will be used; thus, each individual ends up with utility of either one or zero. Therefore, any ‘reasonable’ outcome based rule chooses the voting system in which, in expectation, most people end up being successful.

In front of the veil of ignorance, a utility maximizing agent chooses the voting system that

⁶⁸In the experiment, only one system is selected for payment. It is thus best to always choose the voting system that has the highest expected payoff; risk aversion does not play a role and hedging is not possible.

maximizes the expected utility of any member of the group that this agent will be in. Thus, an individual who knows that she will be in group j chooses the voting system according to

$$(5.3) \quad \mathcal{V}_{F_j, max} = \arg \max_{\mathcal{W} \in \mathbf{W}} E(u_j | \mathcal{W}) = \arg \max_{\mathcal{W} \in \mathbf{W}} p(\text{'individual of } g_j \text{ successful'} | \mathcal{W}),$$

where u_j denotes the utility of an individual of group j after the voting process has been conducted. Here, even if only one issue is voted on, the choice of a utility maximizer in general does not coincide with the choices of individuals with different outcome-based preferences or with the choice of a utilitarian social welfare maximizer.

5.3 Experimental Design and Procedures

The experiment was conducted at the CREED laboratory at the University of Amsterdam with a total of 223 subjects recruited from the CREED subject pool. Participants were primarily undergraduate students, slightly less than half were female, and approximately 60% were majoring in economics or business. The experiment was programmed in PHP/MySQL. Four sessions were conducted, one for each of four treatments. Every participant received 12 euros independently of the choices and outcomes of the experiment. During the experiment, 'points' were used as currency. These points were exchanged for euros at the end of each session at an exchange rate of 1 euro per 80 points. The experiment lasted between 60 and 90 minutes, and participants earned, on average, approximately 17.20 euros. Before starting, the participants had to answer control questions to make sure that they understood the instructions. The experiment did not begin until all participants had successfully answered these questions. Subjects received no information during the experiment on the choices of other subjects. Appendix 5.A provides the instructions and test questions.⁶⁹

5.3.1 Illustration of Voting Systems

In the experiment, subjects choose between different voting systems. These voting systems primarily represent the rules described in the previous section. Subjects are not made familiar with the theories underlying these voting systems. They do not choose between theoretical concepts but between actual voting systems in specified situations. A voting system in a

⁶⁹A pilot with 17 subjects was conducted shortly before the regular sessions were run. The pilot was very similar to the actual sessions but involved some robot players as it was conducted with very few subjects. After the pilot, the instructions and the exchange rate were adjusted.

The experimental sessions consisted of two parts. Subjects received no information regarding the second part before the first part was completed. This chapter is only concerned with the first and main part of the experiment. The sessions including both parts lasted approximately 30 minutes longer than reported, and subjects earned on average 5.30 euros more. More information is available on request.

fixed environment is fully determined by the set of winning coalitions (i.e. when it is known which combinations of votes can pass a proposal). When making their choices, subjects only see neutral graphical representations of the sets of winning coalitions. Subjects' choices thus cannot depend on possible motivations for the specific rules or on whether subjects grasp the concepts underlying these rules.

Figure 5.1 shows a screenshot of a decision situation in the experiment in which two voting systems – large rectangles – are shown (this figure shows the ‘smallest’ voting systems, i.e. the smallest rectangles, used in the experiment). Here, there are four groups: green, red, blue, and yellow. The number of individuals per group is indicated by the number of circles. Thus, the green group has 19 members, the red group has 15, the blue group has 3, and the yellow group also has 3. I now use the voting system on the left side as example. Each row represents a winning coalition. Thus, the first row indicates that if the green, red, and blue groups vote in favor of a proposal, the proposal will be adopted; the second row indicates that if green, red, and yellow vote in favor of a proposal, it will be adopted; and the third row indicates that if green, blue, and yellow vote in favor of a proposal, it will be adopted. The rows shown are all the winning coalitions, except for the grand coalition (everyone voting alike) is obviously always successful and is never shown. In the left voting system of this figure, as a further example, if only the red and the blue groups vote in favor of the adoption of the proposal, the proposal will not be adopted; there is no row showing only the red and the blue group.

5.3.2 Treatments and Overview

The experiment uses a setting of take-it-or-leave-it voting, meaning that the assembly of representatives votes directly on proposals.⁷⁰ Depending on the outcomes of a random draw, group members prefer the proposal to be adopted or rejected (the law governing the random draw depends on the treatment). The set-up of the experiment and its incentives leave no room for strategic voting considerations or abstentions.

The design is primarily a 2×2 factorial between subject design. Subjects are either behind the veil of ignorance (i.e., when they make a choice they do not know which group they will be in if this choice is selected for payment) or in front of the veil of ignorance (i.e., they do know which group they will be in). The other dimension of the 2×2 design determines how preferences over the final outcome are formed. While the ex ante probability of favoring the adoption of the proposal is always one-half, these outcome preferences are either drawn

⁷⁰Most of the literature refers to the adoption or rejection of a proposal. Therefore, I use these terms in this chapter. In the experiment, to avoid leaving subjects wondering about the content of this ‘mysterious’ proposal, the framing used is binary voting on X or Y , where Y is the outcome if no winning coalition of X -supporters can be formed.

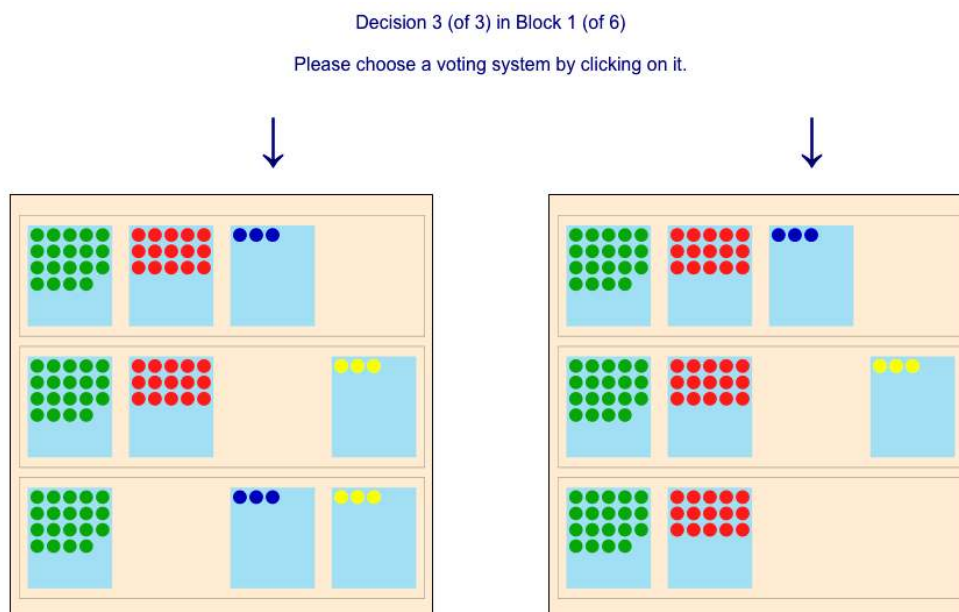


Figure 5.1: Screenshot in front of the veil

Notes: This screenshot is from a treatment in front of the veil (there is an arrow indicating which group the participant will be in). The rows of the graphs are exhaustive lists of winning coalitions, omitting the grand coalition. The voting systems shown here are the ‘smallest’ ones used in the experiment (four groups and four winning coalitions, three without the grand coalition). The voting system on the left is designed according to Rule I, the one on the right according to Rule II.

independently for each participant or they are drawn independently at the group level (in which case these preferences are fully aligned within a group). Table 5.1 summarizes the design.

This is the simplest overview of the design. There are also features that vary within subjects; in each treatment, subjects are shown six different blocks of decision situations (with 18 choices overall). Furthermore, when looking at the treatments in front of the veil, one can split the data according to the group to which subjects belong. Note that it is important to have multiple decision situations in order to be able to identify the theoretical social choice rules.⁷¹

The number of groups (either four or five) and the number of individuals per group are fixed within a block. In each block, there is one voting system representing Rule I (Banzhaf index proportional to the square root of the group size, i.e. Penrose’s Square Root Rule), one voting system representing Rule II (Shapley-Shubik index proportional to the group

⁷¹For each single decision situation one can describe a voting system in many different ways, for example by saying ‘the voting system is such that the majority of groups can pass a proposal as long as the largest group is part of this majority’. Having multiple different decision situations makes it possible to isolate the theoretical rules underlying this research from such ‘ad hoc’ alternatives.

Table 5.1: 2×2 between subjects design

	Independent preferences	Aligned preferences
Behind the veil	<i>BI</i> (54)	<i>BA</i> (54)
In front of the veil	<i>FI</i> (58)	<i>FA</i> (57)

Notes: The cells show the acronyms used for the between subjects treatments (and the numbers of observations). Subjects are either behind the veil of ignorance (*B*) or in front of the veil (*F*). Voting outcome preferences are either drawn independently for each participant (*I*) or are drawn independently at the group level and thus aligned within a group (*A*).

size), and one competing voting system that is different but not determined by any particular rule (called ‘competitor’ as it ‘competes’ with the voting systems designed according to established rules). Subjects always choose between two voting systems. Thus, there are three decisions per block (Rule I - Rule II, Rule I - competitor, Rule II - competitor). Any voting system that a participant chooses may be subsequently selected for payment. The order in which the blocks appear is random to avoid order effects.⁷² Furthermore, the order of the three comparisons within each block is random, as is which voting system appears on the left side of the screen and which appears on the right. Figures 5.1 and 5.2 show screenshots of the decision situations.

The screenshot in Figure 5.1 is taken from a treatment in front of the veil. The participant can thus see an arrow indicating which group she will be in if her choice is selected for payment. In the example shown, she will be in the blue group. The screenshot in Figure 5.2 is taken from a behind-the-veil treatment; thus, there is no arrow indicating which group the participant will be in. The screenshot is from the block with the most complex decisions, i.e., with the largest rectangles: five groups and eight winning coalitions (seven without the grand coalition). I give more detail on all decision situations used in the experiment in Section 5.3.5.

5.3.3 Voting Procedures and Payments

At the end of the experiment, one of the decisions of one participant is selected for payment. The participants are then distributed over the groups involved in the selected decision situation. It is equally likely for each participant to be any of the individuals.⁷³ In the treatments in front of the veil, the participant whose choice is selected for payment is in the group that

⁷²Although the a priori probability that each block is shown at any of the six positions is equal, the first three and last three blocks shown to a participant always have the same number of groups to avoid complicating the situation for the participants.

⁷³The decision situations do not require equal numbers of participants. Some subjects are thus not part of any of the groups and do not receive any payment.

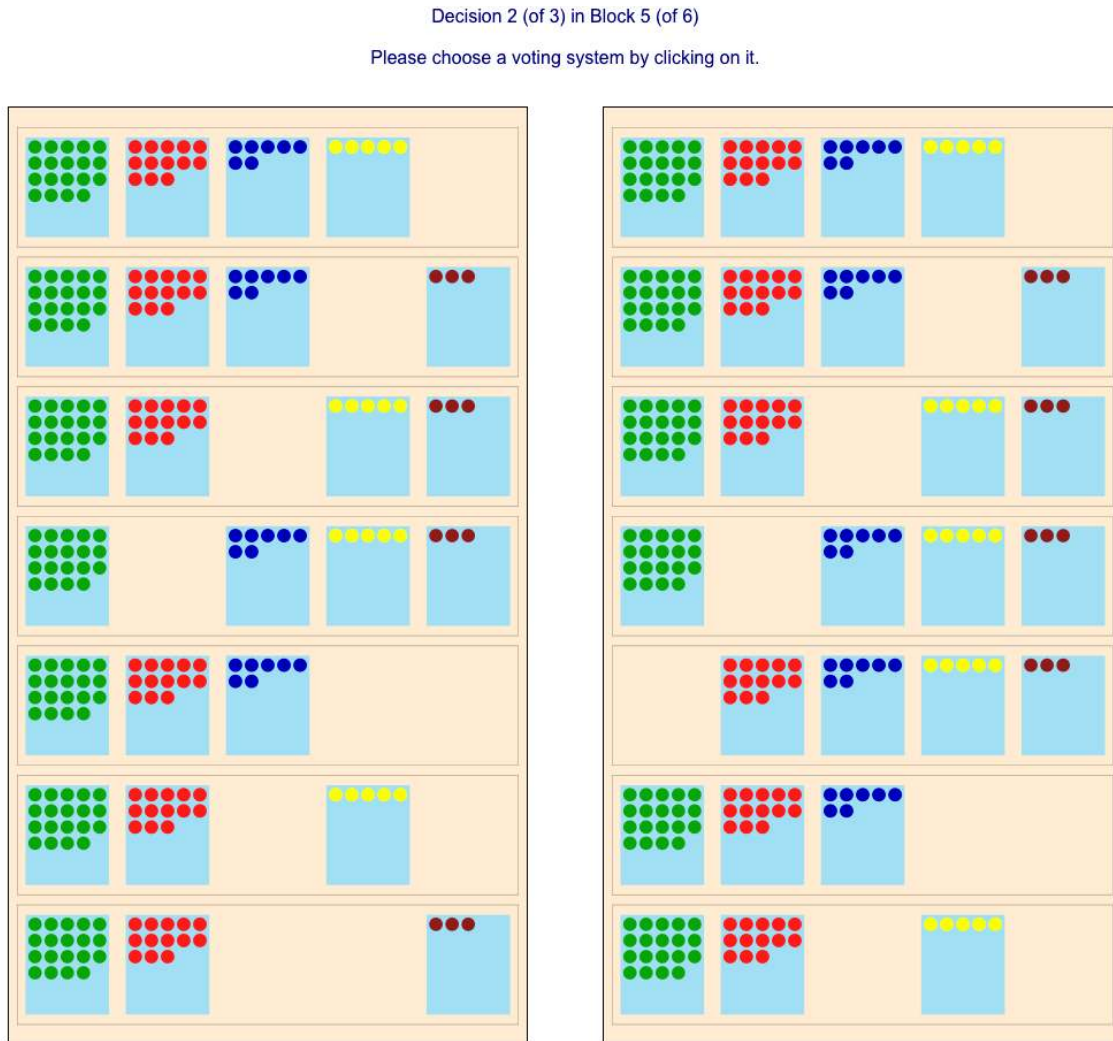


Figure 5.2: Screenshot behind the veil

Notes: This screenshot is from a behind-the-veil treatment (no arrow indicating which group the participant will be in). This decision block is the most ‘complex’ one, i.e., the voting systems shown here are the ‘largest’ ones (five groups, eight winning coalitions – seven without the grand coalition). The voting system on the left is designed according to Rule II, the one on the right according to Rule I.

was indicated to her in the selected decision situation by an arrow. After the participants have been distributed over the groups, their preferences over the outcome are determined. As noted above, ex ante, it is equally likely for each participant to favor the adoption or the rejection of the proposal. In treatments *BI* and *FI*, these outcome preferences are randomly drawn for each participant independently of all other participants. In treatments *BA* and *FA*, the preferences are aligned within a group (i.e., either all members of a group favor the adoption of a proposal or all favor its rejection).

Then, the voting procedure takes place. This voting procedure is fully automated. The

(computerized) representative of each group votes in the best interest of its group. This means that in the treatments with independent outcome preferences, the representative votes according to the outcome preference of the majority of the group; in the other treatments, the representative votes according to the unique outcome preference of all group members. All groups have an odd number of members so that ties are not possible. Whether the proposal is adopted or rejected then depends on the votes of the representatives and on the voting system in place. Note that two assumptions usually made in the literature are fully automated: the assumption that the majority decides at the group level and the assumption that the representative of a group adheres to her group's decision (these two assumptions can be collapsed into saying that 'the representative acts in the best interest of her group'). Then, each member of all groups is paid according to the following rule: if the overall outcome of the voting procedure coincides with the outcome preference of the participant, the participant receives 1000 points; otherwise, the participant receives nothing.⁷⁴

5.3.4 Relationship between Theory and Treatments

The experimental set-up is such that Rule I obtains its best shot in treatment *BI*. There, all assumptions made in the derivation of this rule are satisfied (and the most natural extension for payments is chosen). The treatment is less favorable to Rule II. One cannot really talk about an intensity of choice; thus, it is questionable whether it is reasonable to focus on the pivotal position as described in Section 5.2.1.⁷⁵ One could thus say that Rule I is somehow given an advantage over Rule II in treatment *BI* (because all of its assumptions have been implemented, whereas they are not completely fulfilled for Rule II).⁷⁶ Under group aligned preferences, the assumptions of neither rule are totally fulfilled. This treatment is nevertheless interesting and important; keep in mind that these rules have been proposed many times as solutions for real-world problems in which the assumptions are very far from being fulfilled. It is thus also of interest to see which rule is chosen more often when the assumptions are relaxed (which allows some inferences to be drawn regarding the extent to which the rules correspond to some intuitive concept of fairness/optimality of voting systems). Furthermore, treatment *BA* brings efficiency differences into play, which are absent in treatment

⁷⁴The payments for the experiment are designed in this way so that hedging is not possible for participants and that predictions of utility maximization and other outcome-based preferences coincide (see Section 5.2.2).

⁷⁵Rule II can also be motivated in a bargaining committee setting, which is different from the settings of the experiment and has no direct implication for it. This motivation might play a role if it is somehow connected to people's intuitive feeling of fairness in two-tier voting; but as such, it does not conflict with any of the conclusions of this chapter.

⁷⁶It would also be possible to design an experiment in which Rule II is given its best shot. However, as will be shown below, already in this experiment Rule II is chosen much more often than Rule I. Thus, the additional insights gained from such an additional experiment would probably be limited.

BI. Next, the competing voting systems that do not follow any reasonable rule can be used to check whether the voting systems according to one of the rules have real support in treatment *BI* and whether they are chosen more often than those without any foundation. Finally, the treatments in front of the veil serve as a control to determine whether and how choices differ when subjects know which group they will be in.

5.3.5 Decision Situations

The different voting systems and environments used in the experiment are shown in Table 5.2. The table shows the voting systems as the sets of winning coalitions, which correspond to Rule I, Rule II, and a competing voting system that does not follow any particular rule (I use letters for the different groups in the table instead of colors which are used in the graphical representation in the experiment). As in the graphical illustrations, the grand coalition is always omitted. The table also shows the number of groups and the number of individuals per group for each of the six decision blocks.

Table 5.3 shows the efficiency of the voting systems in the treatment *BA*, i.e. the probability of being successful (having a (computer-determined) preference over the outcome of the voting procedure that coincides with the actual outcome) for an individual behind the veil. The efficiency in treatment *BI* is not shown because it is extremely similar for all voting systems in a block.⁷⁷ More properties of these decision situations and the voting systems, such as efficiency for the treatments in front of the veil and power indices, can be found in Appendix 5.B.1.

It is not trivial to find voting systems that correspond closely to the normative rules used. In Chapter 4, I have stated methods to address this inverse power problem. All voting systems corresponding to one of the rules in this experiment correspond to this rule according to both the methods, based on $err_{group,imp}$ and cv (and therewith also according to the one based on $err_{indirect}$). Thus, even if one of these methods is rejected, the conclusions from the experiment do not lose their validity. To find suitable decision situations, a computer program reviewed (a subset of) all admissible voting systems for all types of possible group compositions. I briefly explain the selection procedure here; further information can be found in Appendix 5.B.2.

To select the decision situations, only groups with an odd number of members were considered to avoid ties within groups during the election of the representative. Furthermore, the groups cannot be too large to fit into the laboratory (the CREED laboratory is large

⁷⁷In the experiment, efficiency differences in treatment *BI* are usually much less than one percent and never much more. Simulations were used to arrive at all efficiency values. For each, the voting procedure (including the preference formation) was simulated two million times.

Table 5.2: Voting systems in the decision blocks

Block #	Groups (size)	Sets of winning coalitions		
		Rule I	Rule II	competitor
1	A(19), B(15), C(3), D(3)	{A,B,C}, {A,B,D}, {A,C,D}	{A,B,C}, {A,B,D}, {A,B}	{A,B,C}, {A,B,D}, {B,C,D}
2	A(21), B(7), C(5), D(3)	{A,B,C}, {A,B,D}, {A,C,D}, {B,C,D}, {A,B}	{A,B,C}, {A,B,D}, {A,C,D}, {A,B}, {A,C}	{A,B,C}, {A,B,D}, {A,C,D}, {A,C}, {A,D}
3	A(27), B(9), C(5), D(3)	{A,B,C}, {A,B,D}, {A,C,D}, {B,C,D}, {A,B}, {A,C}	{A,B,C}, {A,B,D}, {A,C,D}, {A,B}, {A,C}, {A,D}	{A,B,C}, {A,B,D}, {A,C,D}, {B,C,D}, {A,C}, {A,D}
4	A(15), B(13), C(11), D(5), E(1)	{A,B,C,D}, {A,B,C,E}, {A,B,D,E}	{A,B,C,D}, {A,B,C,E}, {A,B,C}	{A,B,C,D}, {A,C,D,E}, {A,C,D}
5	A(17), B(15), C(7), D(5), E(5)	{A,B,C,D}, {A,B,C,E}, {A,B,D,E}, {A,C,D,E}	{A,B,C,D}, {A,B,C,E}, {A,B,D,E}, {A,B,C}	{A,B,C,E}, {A,C,D,E}, {B,C,D,E}, {B,C,E}
6	A(19), B(13), C(7), D(5), E(3)	{A,B,C,D}, {A,B,C,E}, {A,B,D,E}, {A,C,D,E}, {B,C,D,E}, {A,B,C}, {A,B,D}	{A,B,C,D}, {A,B,C,E}, {A,B,D,E}, {A,C,D,E}, {A,B,C}, {A,B,D}, {A,B,E}	{A,B,C,D}, {A,B,C,E}, {A,B,D,E}, {A,B,C}, {A,B,D}, {A,B,E}, {A,B}

Notes: For each of the six decision blocks, the table shows the groups, their sizes, and the sets of winning coalitions (the voting systems) according to Rule I, Rule II, and a competitor that does not follow any particular rule. The grand coalition is always omitted. More details on the different voting systems and decision blocks can be found in Appendix 5.B.1.

enough to handle up to 58 subjects simultaneously). All constellations have either four or five different groups. The number of groups and the number of individuals per group are constant in a decision block, as is the number of winning coalitions (the voting system that corresponds best to a rule is thus always selected out of all admissible voting systems with

Table 5.3: Efficiency of the voting systems in treatment *BA*

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
Rule I	0.684	0.722	0.756	0.601	0.616	0.663
Rule II	0.713	0.760	0.793	0.608	0.622	0.673
competitor	0.672	0.747	0.739	0.586	0.581	0.670

Notes: This table shows the efficiency, i.e., the probability of success for an individual behind the veil, of the voting systems when outcome preferences are aligned within groups.

a fixed number of winning coalitions).⁷⁸ Holding the number of winning coalitions constant makes it impossible for participants to choose according to the simple but good heuristic of always taking the system with the most winning coalitions. All of the different decision situations were selected in such a way that the recommended voting systems according to either rule are the same no matter how the inverse power problem is addressed. These two voting systems are furthermore different from each other and such that each system does not perform very well in terms of the other rule. As mentioned before, a competitor, i.e. a voting system that does not perform well according to either rule, is added.

5.4 Results

First, I present results on the choice of voting systems behind the veil of ignorance. Then, I examine whether subjects' decisions behind the veil can be explained by utility maximization or other outcome-based behavior. Finally, I investigate whether being behind or in front of the veil of ignorance makes a difference in this experiment and what drives choices in front of the veil. I present the results in this way to make it easy to understand the main findings of this chapter.

Subjects in the experiment receive no information on others' decisions; thus, observations can be treated as statistically independent. Only non-parametric tests are shown because these draw upon less restrictive assumptions concerning underlying distributions than parametric tests. All tests performed are two-sided. The number of observations is as shown in Table 5.1. Additional graphs and data can be found in Appendix 5.B.3.⁷⁹

⁷⁸The number of winning coalitions has sometimes been called 'efficiency' in the literature, going back to Coleman's 'power of a collectivity to act' (Coleman, 1971). Referring to the number of winning coalitions as 'efficiency' is avoided in this chapter.

⁷⁹There are no particularly interesting gender effects, neither do nationality or field of studies have a strong influence on the results. Therefore I do not present the data split according to any of these attributes. The experiment is quite complex for subjects. Because this complexity could be foreseen before the experiment was conducted, the questionnaire contained a question asking how choices were made, with the possibility of answering that the choices were made 'more or less randomly'. I use the full data set in general, but at points, I also refer to the data excluding subjects who chose randomly (these 'restricted' data do not contradict the

5.4.1 Subjects Prefer Rule II (Proportional Shapley-Shubik Power) over Rule I (Penrose's Square Root Rule)

The primary research question of this chapter is the type of voting systems subjects choose behind the veil of ignorance. Do they prefer voting systems according to Rule I, according to Rule II, or do they not consistently choose according to either of these two rules?

A good way to summarize the data of the behind-the-veil treatments is to examine how many participants predominantly choose one voting system. Each participant makes 18 choices overall, and each of the three types of voting systems is involved in 12 of these choices. Figure 5.3 shows how many participants in treatments *BI* and *BA* choose a particular system at least 9 out of 12 times (considering 10 or 11 out of 12 choices provides a similar picture; this can be seen in Appendix 5.B.3, Figure 5.23). One can see that in both treatments, there are many more subjects who overwhelmingly choose the Rule II voting system than subjects who overwhelmingly choose the Rule I voting system. Choices for the competitor are assumed to primarily reflect noisy behavior.

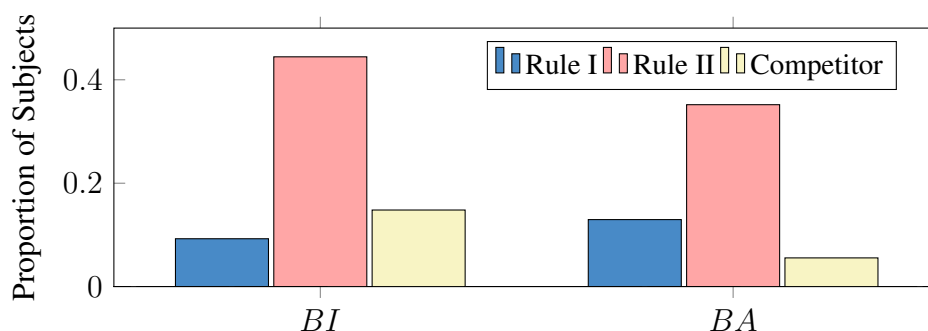


Figure 5.3: Proportion of participants predominantly choosing one system

Notes: The figure shows the proportion of participants who choose a type of voting system at least 9 out of 12 times in treatments *BI* and *BA*.

Figure 5.4 shows participants' choices in more detail. Because there are six blocks, each participant chooses between voting systems according to two particular 'rules' (i.e. the two social choice rules and the competitor) six times. Now, we consider how often one of the rules is preferred. This yields for each participant and each comparison between two rules a number between 0 and 6, where 3 means that each rule was chosen equally often in the direct comparison (0 means that the first mentioned rule – Rule I in 'R1-R2' and 'R1-c', Rule II in 'R2-c' – was never chosen over the second mentioned rule; 6 means that the first mentioned rule was always chosen). Figure 5.4 shows a bar for each participant and for each direct

results of the full data). The data without participants who answered that they chose 'more or less randomly' contain fewer observations, namely, 35 (*BI*), 32 (*BA*), 40 (*FI*), and 35 (*FA*).

comparison between two rules, indicating how often each of the rules was chosen over the other.⁸⁰

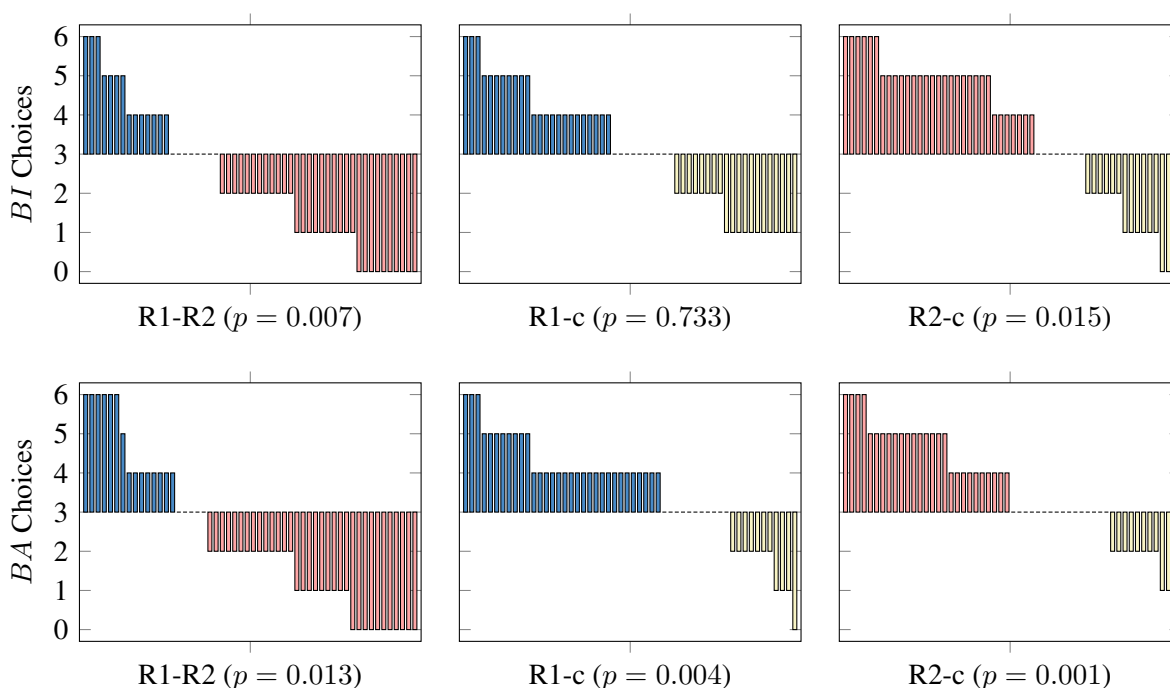


Figure 5.4: Participants' choices

Notes: Each bar corresponds to a participant. The values are between 0 and 6. In the comparison, 'R1-R2' 0 means that a participant has chosen Rule I voting systems zero times in comparison with Rule II voting systems (and thus has chosen Rule II systems six times). The p -values stem from two-sided Wilcoxon signed rank tests.

I use Wilcoxon signed-rank tests to determine whether the differences in choices are statistically significant. The p -values are shown in the figure. The null hypothesis of each test is that there is no difference in how often the voting systems were chosen. Thus, the voting systems of Rule II are chosen significantly more often than the systems of Rule I in both treatments, *BI* and *BA*. Rule II voting systems are also chosen significantly more often in both treatments than the competing voting systems not following any particular rule. Rule I voting systems are chosen significantly more often than the voting systems not following any particular rule only in treatment *BA*. Thus, the following result is obtained:

Result 1. *Participants behind the veil of ignorance prefer voting systems according to Rule II over voting systems according to Rule I. This preference exists both when efficiency concerns are absent and when such concerns are present.*

⁸⁰Means (medians) of these comparisons are as follows. BI: 2.23(2), 3.11(3), 3.63(4); BA: 2.31(2), 3.56(4), 3.63(3.5); in the order 'R1-R2', 'R1-c', 'R2-c'.

As explained in Section 5.3.4, the cleanest scenario to test the preference between the two rules is treatment *BI*. In this treatment, efficiency concerns are absent because all systems are basically equally efficient and the assumptions for the derivation of the rules are relatively well satisfied (perfectly for Rule I, a bit less so for Rule II). We can thus see that subjects prefer Rule II voting systems over Rule I voting systems when there is no expected payoff difference. Furthermore, although Rule I is given its best shot in *BI*, subjects do not even choose voting systems according to it significantly more often than voting systems not corresponding to any specific rule. Comparing the results in treatments *BI* and *BA* shows that subjects' choices are relatively robust to changes in the correlation structure of voting outcome preferences; although the outcomes might not be equal, the patterns are very similar. The results in treatment *BI* cannot be explained by expected utility maximization alone.⁸¹ The next section explores whether utility maximization plays a role in participants' choices behind the veil in general.

5.4.2 Subjects React to Changes in the Payoff Structure

Result 1, as far as treatment *BI* is concerned, cannot be explained by expected utility maximization alone. Next, I consider whether subjects in treatment *BA* choose the voting systems predicted by utility maximization (which are also predicted by other outcome based preferences, as explained in Section 5.2.2). Figure 5.5 depicts each of the 18 choice situations in treatment *BA*. The first three choices are the choices of block 1, the next three of block 2, and so on. Within each block, the first choice bar corresponds to the proportion of Rule I systems chosen over Rule II systems, the second corresponds to Rule I systems versus the competitor, and the third choice bar corresponds to Rule II systems versus the competitor. The choice bars show how much more often one system was chosen than the other (for example, a value of -0.15 for a bar means that the first voting system was chosen 15 percentage points less often than the second). The payoff bars represent the difference in expected payoffs between the two systems concerned. They have been scaled such that the sum of absolute values is equal to the sum of absolute values of the choice bars (i.e., the total area of both types of bars is equal).

Noise-free perfect utility maximization would thus mean that whenever a payoff bar is positive, the corresponding choice bar should be at plus one, and when a payoff bar is negative, the corresponding choice bar should be at negative one. Less, extremely, it is interesting to see whether subjects consistently choose the more efficient system more often (i.e., whether the choice bars are of the same sign as the corresponding payoff bars). This is indeed

⁸¹Of course, because there are basically no differences in expected payoffs, any choice can be rationalized. However, utility maximization alone cannot explain systematic differences as observed in treatment *BI*.

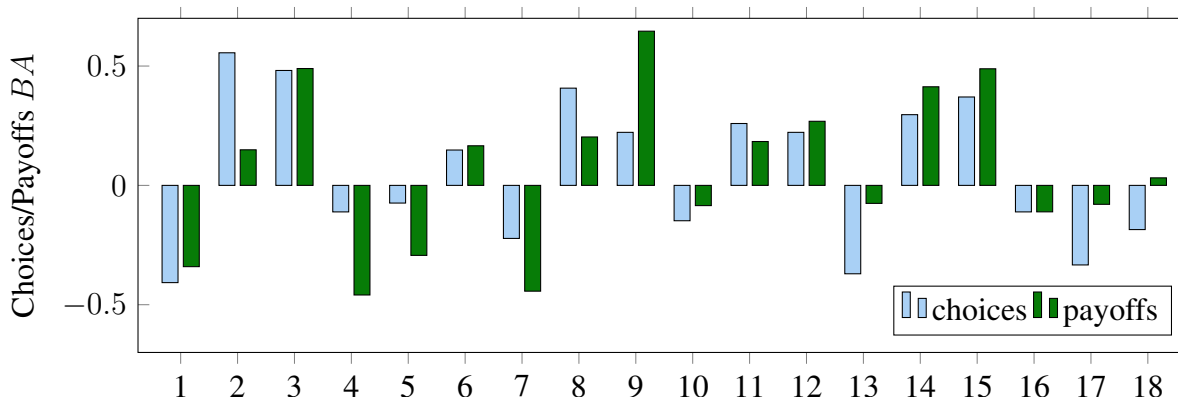


Figure 5.5: Differences in choices and expected payoffs in treatment *BA*

Notes: This graph shows differences in choices and expected payoffs for all 18 choice situations in *BA*. The first three choices are from block 1 and so on. Within each block, the first choice bar corresponds to ‘R1-R2’, the second corresponds to ‘R1-c’, and the third corresponds to ‘R2-c’. The y-axis shows how much more often one system was chosen. The payoff bars represent the respective difference in expected payoffs, rescaled so that the total area of both types of bars is equal.

the case. The correlation between the differences in choices and the differences in expected payoffs is 0.756. A Pearson’s product-moment correlation test yields a p -value of less than 10^{-3} , rejecting the null hypothesis of zero correlation.

There is thus a positive correlation between the voting systems that subjects choose in treatment *BA* and the expected payoffs of these voting systems. This correlation may stem from different causes, however. For example, it could be the case that the most efficient voting systems usually coincide with Rule II voting systems. Indeed, Rule II voting systems are more efficient than Rule I voting systems (and Rule I tends to be more efficient than the competitor). With a difference-in-differences analysis considering how the outcomes in *BI* and *BA* differ, one can correct for these ‘general preferences’ (i.e. the preferences in the absence of expected payoff differences).

For this purpose, I consider the correlation between the differences in choices between *BA* and *BI* and the differences in payoffs between *BA* and *BI*.⁸² If this correlation is positive, it means that if people choose a voting system relatively more often in *BA* than in *BI*, on average, this goes together with an increase in expected payoff. Figure 5.6 shows the correlations of choices and expected payoffs in the difference-in-differences version.

This correlation is indeed positive, but at 0.193, it is relatively low and statistically not

⁸²To be precise, for each decision situation – say, between voting systems X and Y – I consider the correlation between ‘percentage point difference of voting system of type X chosen versus voting system of type Y in treatment *BA* minus percentage point difference of voting system of type X chosen versus voting system of type Y in treatment *BI*’ and ‘expected payoff of voting system X in *BA* minus expected payoff of voting system Y in *BA* minus [expected payoff of voting system X in *BI* minus expected payoff of voting system Y in *BI*]’. Of course, because payoff differences in *BI* are basically zero, the part in brackets is always very close to zero.

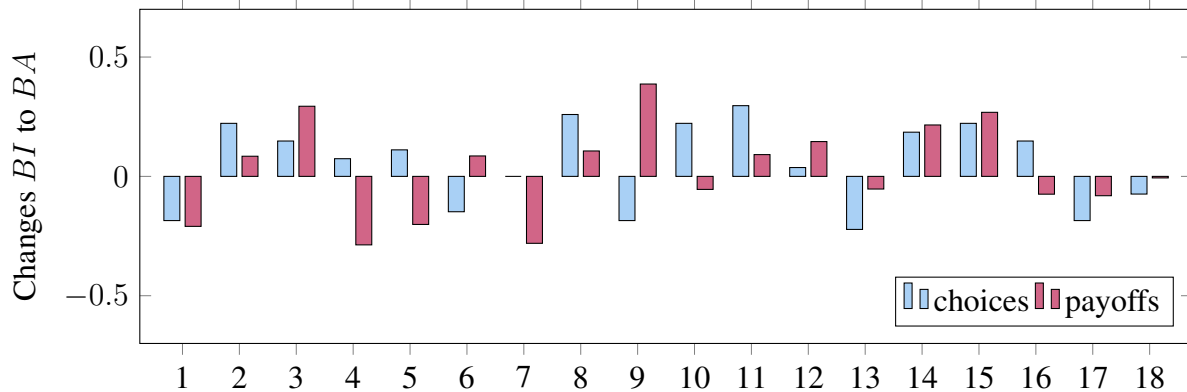


Figure 5.6: Difference in differences BI to BA , choices, and expected payoffs

Notes: The graph shows the difference in differences from treatment BI to BA in choices and expected payoffs in the 18 different choice situations (in the same order as for Figure 5.5). Expected payoff differences are scaled so that the total area of payoff difference bars and choice difference bars is equal.

significantly different from zero (the p -value of a Pearson's product-moment correlation test is 0.443). Taking the data without the observations of subjects who stated that they chose more or less randomly (see Footnote 79), the correlation increases to 0.661, and the difference from zero is statistically significant with a p -value of 0.003. Thus, there is evidence that subjects take payoff/utility considerations into account when making their choices in treatment BA where efficiency differences are present. Keep in mind, however, that the choices across the treatments BI and BA are quite similar (as shown in Figure 5.4) and that I investigate here whether subjects choose more in line with the predictions by utility maximization and other outcomes based preferences than in treatment BI . There is thus the following result (keeping in mind that the predictions of utility maximizing behavior coincide with the predictions of basically all other outcome-based preferences behind the veil of ignorance):

Result 2. *In treatment BA with differences in expected payoffs, participants choose voting systems that give them higher expected payoffs more often than in treatment BI without differences in expected payoffs.*

5.4.3 Subjects Choose Differently When They Are in Front of the Veil (to Their Own Group's Benefit)

Does it matter whether subjects know which group they will be in? Or are subjects in the laboratory so selfless (or confused) that they choose the same no matter which group they will be in? The answer is that in front of the veil of ignorance, subjects overwhelmingly choose in the interest of their own group. Note that here, voting power considerations and payoff considerations generally lead to the same outcome: the voting systems that are beneficial for

one group have generally high voting power for this group (according to the Banzhaf index as well as the Shapley-Shubik index) and high expected payoffs.

Figure 5.7 shows the choices of participants in treatment FA who are either in the smallest or in the largest group. These choices are shown for all three comparisons in all six blocks. One can see that subjects in the small groups choose very differently from subjects in the large groups in each of the six blocks. Usually, when subjects in the smallest group favor one voting system over another, subjects in the largest group favor the other voting system. The voting systems that subjects of a group prefer are, in general, those that give more power and greater expected payoff to their group. For example, considering the fifth block, subjects in the smallest group in treatment FA prefer Rule I systems over Rule II systems and the competitor over both Rule I and Rule II systems. Indeed, this ordering is best for their group. The Banzhaf voting powers of this group for the three different voting systems are 0.176, 0.067, and 0.333 (in the order: Rule I, Rule II, competitor); the Shapley-Shubik powers are 0.15, 0.05, and 0.383; and the probabilities of success are 0.594, 0.532, and 0.657, respectively. The largest group chooses the Rule I and Rule II voting systems much more often than the competitor. Indeed, voting power and expected payoffs are considerably higher for these systems for the large group. Differences between the Rule I and Rule II systems in this block are negligible for the large group; thus, it comes as no surprise that one system is only chosen slightly more often than the other. In general, this pattern holds roughly across all blocks and groups and similarly for treatment FI : subjects choose voting systems that are good for the group they are in according to both voting power and expected payoff. The respective data can be found in Appendix 5.B.3.

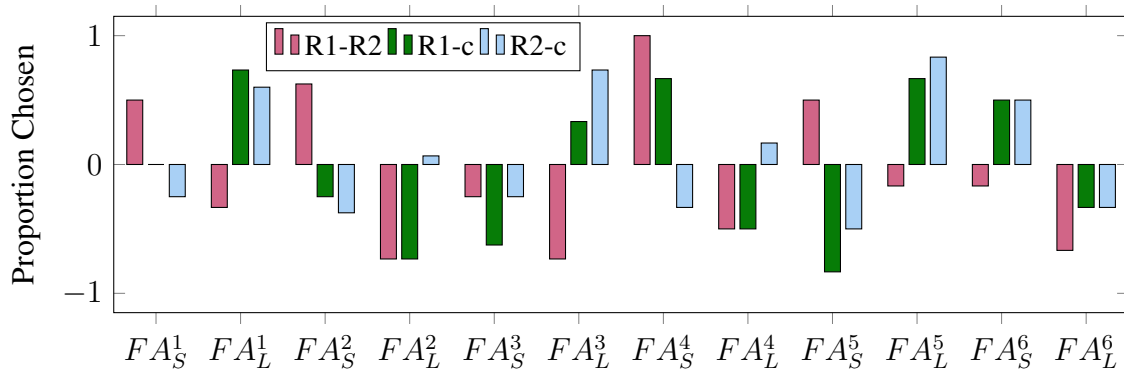


Figure 5.7: Choices in front of the veil for the smallest and largest groups

Notes: The figure shows choices in treatment FA . The subscripts S and L denote the smallest and largest groups, respectively. The superscripts represent the decision block. The bars ‘R1-R2’ show how often voting systems according to Rule I have been chosen over voting systems according to Rule II (similar for ‘R1-c’ and ‘R2-c’).

Next, I test whether subject’s choices are significantly different behind and in front of

the veil of ignorance, where the data in front of the veil are split according to group membership. For this purpose, I use the data on which system is chosen over both other systems in each decision block and Fisher's exact test. The null hypothesis of this test is that there is no difference in the proportions of choices between participants in different categories (i.e. treatments; the data can be found in Appendix 5.B.3, Table 5.12). For independent voting outcome preferences, the different categories for each decision block are BI , FI_A , FI_B , FI_C , FI_D and for blocks 4 to 6 also FI_E (the subscripts represent the different groups in front of the veil); similarly for aligned voting outcome preferences. Table 5.4 shows the p -values of these tests per block and for independent and aligned outcome preferences separately.

Table 5.4: Differences in choices behind and in front of the veil

Block	p -values		Holm-Bonferroni p -values	
	independent pref.	aligned pref.	independent pref.	aligned pref.
1	0.045	0.034	0.089	0.034
2	0.295	$< 10^{-3}$	0.295	0.003
3	$< 10^{-4}$	$< 10^{-5}$	$< 10^{-3}$	$< 10^{-4}$
4	$< 10^{-4}$	0.001	$< 10^{-3}$	0.003
5	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	0.003
6	$< 10^{-4}$	$< 10^{-3}$	$< 10^{-3}$	0.002

Notes: This table shows the p -values of Fisher's exact test (null-hypothesis: proportions of preferred voting systems are equal). The categories for each decision block are BI , FI_A , FI_B , FI_C , FI_D and, where applicable, FI_E for independent preferences, similarly for aligned preferences.

The results from these tests are overwhelmingly clear: in almost all blocks, the outcome that subjects choose differently is statistically highly significant. Nevertheless, what we actually want to know is whether there is a *systematic difference in at least one block*; this establishes that choices are different. Therefore, to be completely correct and to address potential problems of multiple testing, I also report Holm-Bonferroni corrected p -values in Table 5.4 (these p -values are naturally larger throughout, but already one significant finding stands for a systematic difference). This leads us to the last result:

Result 3. *Participants' choices are different in front of the veil and behind the veil of ignorance (participants in front of the veil choose to the benefit of their own group).*

5.5 Discussion

The main result of this chapter is that people shy away from voting systems designed according to Penrose's Square Root Rule, even when this rule is given its best shot. This is an important finding. A policy maker, researcher, or anyone with the task to design a voting mechanism may still prefer a voting system designed according to Penrose's Square Root Rule for theoretical reasons. However, this research suggests that the implementation of such a voting system is problematic, because the people who will be subjected to it may not accept it. The acceptance of voting procedures is important for democratic legitimacy reasons and can also influence outcomes through people's actions (as shown by the literature on procedural fairness). Rather than accepting voting systems designed according to Penrose's Square Root Rule, people may be much more willing to accept those following the rule of proportional Shapley-Shubik power.

How can these preferences be explained? I have designed the experiment in a way to be able to exclude the most common behavioral explanations: The general preference for voting systems exhibiting proportional Shapley-Shubik over systems designed according to Penrose's Square Root Rule in the treatment where efficiency concerns are absent cannot be explained by any outcome based concept. Thus, utility maximization, altruism, inequality aversion, social welfare maximization, loss aversion, or risk aversion cannot explain these findings. A possible interpretation is that proportional Shapley-Shubik power corresponds to people's intuitive sense regarding which types of voting systems are good (more specifically, to an intuitive sense about how much power smaller groups and how much power larger groups should have). People like to give larger groups relatively more power than they would get according to Penrose's Square Root Rule.

When looking at the results from the treatment where outcome preferences are aligned within groups, the findings can be interpreted as people starting from their intuitive feeling (predominantly choosing voting systems exhibiting proportional Shapley-Shubik power) and adjusting their choices in the direction that gives them a higher expected payoff. This interpretation is consistent with the choice pattern observed in the treatment with independent outcome preferences, and it is furthermore consistent with the fact that, on the one hand, participants' choices are relatively robust to changes in the correlation structure of outcome preferences and, on the other hand, payoffs can be shown to have an impact on participants' choices.

The last result, showing that participants choose voting systems to their own benefit when they are in front of the veil, gives additional support to the other results. This result shows that it is possible to meaningfully introduce a veil of ignorance in such an experiment (and that subjects understand the payment structure, the voting stage of the experiment, and the

graphical representation of the voting systems). Of course, outside of the laboratory, people are usually in front of the veil of ignorance. Nevertheless, one can assume that the voting systems that are accepted most are the voting systems that are predominantly chosen behind the veil of ignorance.

One possible criticism of many laboratory experiments is that the student subject pool is not representative of the general population. Concerning this particular experiment, it is unlikely that the results depend on the composition of the subject pool. Subjects in front of the veil of ignorance primarily choose to their own benefit, whereas subjects behind the veil predominantly choose voting systems exhibiting proportional Shapley-Shubik power; this treatment difference might be slightly more or less pronounced in the general population, but it is very unlikely that it would be reversed (the same holds for the other results; discussions of the appropriateness of student subject pools in economic experiments can be found in Falk and Fehr, 2003, Falk and Heckman, 2009, and Schram, 2005). Furthermore, these preferences are not particularly different between women and men, Dutch and non-Dutch participants, or students of economics/business and other students, which further supports the view that these preferences are unlikely to be very different in a more representative population.

Appendix 5.A Instructions, Test Questions, and Questionnaire

Here, the instructions and test questions of the experiment can be found. Because a considerable number of graphical illustrations were used, I show screenshots. The first set of instructions corresponds to treatment *BI*. The differences in the instructions between treatments concern only a few screens. After the instructions for *BI*, the screens that are different for the other treatments will be shown. Note that the answers to the test questions can differ depending on the treatment (the screens of the test questions are the same). The questionnaire in the end asks for the following attributes and the following questions for the part of the experiment reported in this chapter. Answer possibilities where present are in parentheses.

- Gender (Male/Female)
- Age
- Have you participated in a CREED experiment before? (No / Yes, once or twice / Yes, more than twice)
- Nationality
- How clear were the instructions of the experiment (Clear / Not as clear as possible, but understandable / Unclear)
- Which of the following comes closest to your field of study [multiple answer possibilities, not reproduced here]
- Which of the following describes your decisions in the first part of the experiment best? (I have tried to make the decisions in a way that I thought was sensible. / I have made my decisions more or less randomly, because I didn't really understand the task and/or its consequences. / I have made my decisions more or less randomly for other reasons (if so, please specify below).)
- If you had a certain way of making decisions in the first part of the experiment, can you describe it very briefly?
- How would you describe your command of English? (Excellent / Very good / Good / Fair / Bad)
- Are there any comments you would like to leave for us?

5.A.1 Screenshots of Instructions and Test Questions, Treatment *BI*

Figures 5.8 to 5.19 contain screenshots of the instructions and test questions for treatment *BI*. They appear in the same order as in the experiment. Whenever screens are not the same in all treatments, the treatments corresponding to the screen are noted in the caption.



Welcome to this experiment!

This experiment is anonymous; the data from your choices will only be linked to your station id, not to your name. You will be paid privately at the end, after all participants have finished the experiment. The experiment will take approximately 2 hours. You will spend a considerable fraction of this time reading instructions.

This experiment consists of two parts. You will receive instructions for the second part before it begins. **Please read all instructions VERY carefully**, otherwise you might be lost later on in the experiment. There will be test questions before you can continue to the experiment. You can use the menu on top of the screen to go back to previous parts of the instructions.

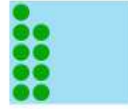
Payments during the experiment are in points. 80 points will be exchanged into 1 Euro. For showing up and answering all questions carefully you receive a fixed payment of 12 euros. Everything you earn during the experiment will come on top of this.

You are not allowed to speak with other participants or to communicate with them in any other way. **If you want to ask a question at any time, please raise your hand and someone will come to your desk.**

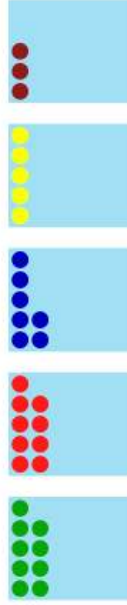
[Next](#)

Figure 5.8: Welcome screen

In the experiment there will be different groups. The groups can have different sizes. A group is represented by a small rectangle with circles in it, each circle stands for one group member. A group of 9 individuals would thus look as shown below (colors have no meaning other than to distinguish different groups).



In the experiment there will always be either 4 or 5 groups. If there are 5 different groups, they could look as shown below (remember, each circle stands for one group member). If there are only four groups, there will be no brown group.



The colors from left to right are always: green, red, blue, yellow, sometimes brown.

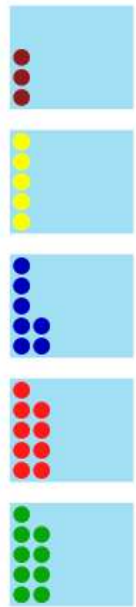
Together, the groups have to vote. There are two possible 'outcomes', called X and Y. Each group has exactly one vote. It can either vote for X or for Y, abstention (i.e. not voting) is not possible. (Note that it is the groups that vote and not the individuals in the groups.)

The final outcome is the result of how each group votes and of the voting system used. Your decisions in today's experiment will determine the voting system. You will thus have to understand very well what such a voting system is and how it works.

[Next](#)

Figure 5.9: Instructions voting 1

Assume that there are the following five groups (the same as shown before).



As mentioned, the groups can vote for either X or Y. There are different combinations of group votes possible, one would for example be: the green, the red, and the yellow groups vote for X, the blue and the brown groups vote for Y. We will illustrate a combination of votes by only showing the groups that vote for X. All other groups then vote for Y.

Thus, our example (green, red, yellow groups vote for X – blue, brown groups vote for Y) would look as shown below. The green, the red, and the yellow groups are shown (the X-votes) and the places where the groups would be located that vote for Y are left empty (the spots for the blue and the brown groups).



[Next](#)

Figure 5.10: Instructions voting 2

A voting system decides for all possible combinations of (group) votes whether the outcome will be X or Y. Thus, you can think of a voting system as a list that contains all possible combinations in which the groups can vote and for each of these combinations it says whether the outcome is X or Y. Instead of giving this whole list it is sufficient to give a list of all combinations of group votes that lead to outcome X. All the combinations that are not in the list then automatically lead to outcome Y.

All in all, we will describe a voting system by providing all combinations of X-voting groups that yield outcome X. For example, if you see a row showing a green and a red group and empty spots where other groups would be, this means that the situation "the green and the red groups vote for X while the other groups vote for Y" leads to outcome X.

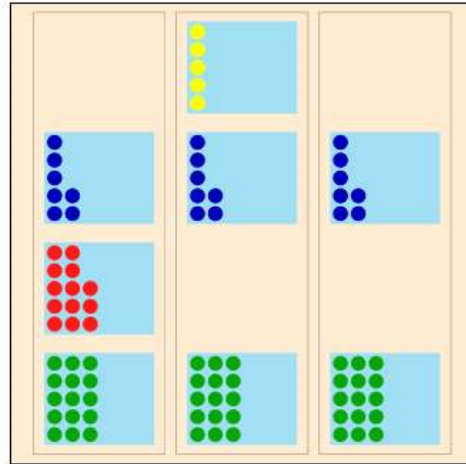
We illustrate this as follows. A voting system is a large rectangle with multiple rows. Each row shows a combination of group votes that leads to the outcome X (remember, within one row, groups that are shown are X-votes, groups that are left out are Y-votes). If a combination of votes is not shown as a row in the large rectangle, this combination does not lead to the outcome X and thus it leads to the outcome Y. There is one exception, however. If all groups vote for X the outcome will always be X – as this is self-evident the row representing all groups voting for X is never shown in the rectangle.

To summarize: For any voting system, if you want to know for a certain combination of votes whether the outcome will be X or Y, you check all rows of the voting system and check whether there is a row representing this combination. If you find such a row the outcome will be X, if you do not find such a row the outcome will be Y.

Take a look at the very simple voting system on the right. There are four groups (there is no brown group).

The first row says that if the green, red, and blue groups vote for X and the yellow group votes for Y, the outcome is X. The second row says that X-votes of green, blue, and yellow lead to the outcome X. And the third one says that the X-votes of green and blue alone lead to outcome X.

Except for the case that all groups vote for X, these combinations of votes are the only ones that lead to outcome X. Thus, for example the situation "red, blue, and yellow vote for X, green votes for Y" leads to outcome Y (there is no row showing only the red, blue, and yellow group). Similarly, if only the green and the yellow groups vote for X, the outcome will be Y.



[Next](#)

Figure 5.11: Instructions voting 3

Additional information: All voting systems that you will see in this experiment have the following property. If a combination of groups voting for X leads to the outcome X, then the outcome will also be X if another group votes for X in addition. This means for example that if you, as below, see a row showing only the green, the red, and the brown group (fifth row), you can be sure that there will also be a row showing the green, the red, the brown, and the blue group (second row) and a row showing the green, the red, the brown, and the yellow group (third row).

Let's look at more examples with the voting system shown on this page.

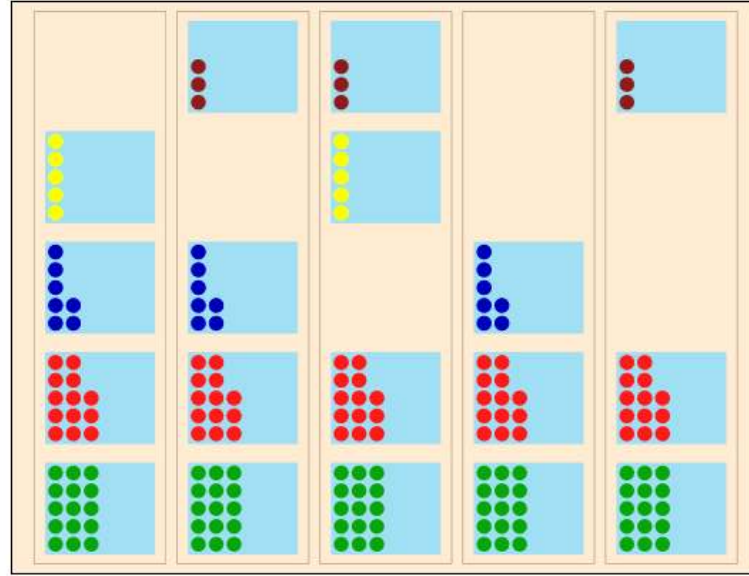
If the green, red, and blue groups vote for X and the yellow and brown groups vote for Y, what will be the outcome? The answer is X, because the fourth row represents this combination of votes (green, red, and blue are shown, i.e. vote for X, yellow and brown are left out, i.e. vote for Y; this row is in the large rectangle, thus the outcome will be X).

If instead the red, blue, yellow, and brown groups vote for X and the green group votes for Y, what will be the outcome? The answer is Y, because this combination is not represented in the rectangle (there is no row without the green and with the red, blue, yellow, and brown groups).

How about if all groups vote for X? The outcome will be X even though there is no row with all groups – as mentioned above this row is always left out, because self-evidently all groups agreeing on X leads to the outcome X.

What if all groups with at least 6 members vote for X and the others vote for Y? The brown and the yellow groups have fewer than 6 members (3 and 5) and thus vote for Y. The green (15), the red (13), and the blue (7) groups have more than 6 members and thus vote for X. This means that the outcome will be X (fourth row).

How about if the green, red, and yellow groups vote for X, the others for Y? The outcome will be Y as there is no row showing just the green, red, and yellow groups. Does this change if instead of the yellow 5-person group the brown 3-person group votes for X? The answer is yes, then the outcome will be X, because the fifth row shows the green, the red, and the brown groups. You can thus see that if you replace an X-vote of a larger group with an X-vote of a smaller group it might sometimes change the outcome from Y to X – it depends on the voting system.



[Next](#)

Figure 5.12: Instructions voting 4

All questions refer to the voting system illustrated on the right.

If the brown group votes for Y, can the outcome ever be X?

- Yes, it depends on the votes of the other groups.
 No.

If the brown group votes for X, can the outcome ever be Y?

- Yes, it depends on the votes of the other groups.
 No.

If three groups vote for X and two groups for Y, what will be the outcome?

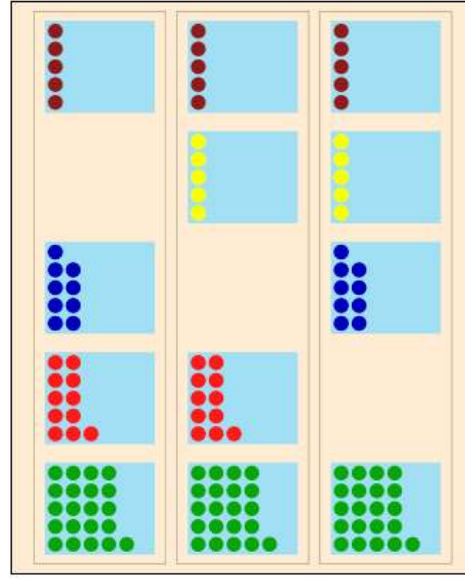
- X
 Y
 That depends which groups vote for X and which for Y.

If the red group votes for X, what will be the outcome?

- X
 Y
 That depends on the votes of the other groups.

If all groups vote for X, what will be the outcome?

- X, it does not matter that there is no row showing all groups.
 Y, because there is no row showing all groups.



Send

Figure 5.13: Test questions voting 1

All questions refer to the voting system illustrated on the right. (Now there are only 4 different groups, green, red, blue, and yellow.)

If the green, the red, the blue, and the yellow group vote for X, what will be the outcome?

- X
- Y

If the group with 13 members and the group with 7 members vote for X, will the outcome always be X?

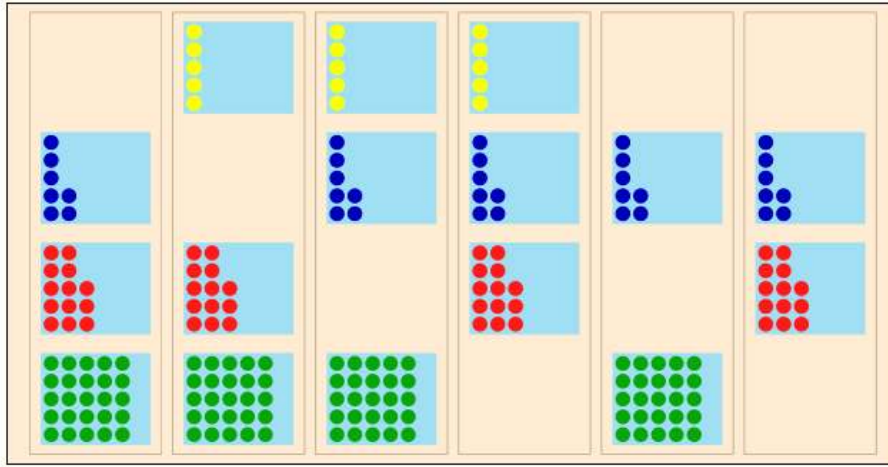
- Yes.
- No, it depends on the votes of the other groups.

If the green and the red group vote for Y, will the outcome always be Y?

- Yes.
- No, it depends on the votes of the other groups.

If the green and the yellow group vote for Y, will the outcome always be Y?

- Yes.
- No, it depends on the votes of the other groups.



Send

Figure 5.14: Test questions voting 2



By now you should have understood very well what a voting system is and how it is illustrated. There will be some more instructions, explaining what you have to do in part 1. Note that you cannot go back to the explanation of the voting system after you click *Continue*.

If anything is unclear raise your hand and someone will come to your desk.

[Continue](#)

Figure 5.15: Continuation screen

The groups vote to choose between the two possible outcomes X and Y. You know now how the outcome is determined given the votes of the groups and given a voting system (the voting system used will depend on your choices). What the groups vote for depends on the preferences of the group members, which are the participants of the experiment.

Each participant either prefers X or Y. Whether you prefer X or Y will not be determined until the end of the experiment. The probability that you prefer X (or Y, similarly) is one half. This probability is the same for all participants independently of the group they are in and independently of the random preferences of other participants.

If the majority of members of a group prefers X the group will vote for X, if the majority of members prefers Y the group will vote for Y (all groups have an odd number of members, ties are thus not possible). This is done automatically, you cannot change the way your group votes.

The outcome (X or Y, determined by the group voting) will determine your earnings. Your earnings are 1000 points if your preference is equal to the outcome. If not, your earnings are 0.

[Next](#)

Figure 5.16: Instructions choices and payoffs 1, treatments *BI* and *FI*



During part 1, you will be asked to make 18 decisions. Each decision consists of choosing one out of two voting systems. You will see a screen with one voting system on the left and one on the right. You choose one of these voting systems by clicking on it. In the end of the experiment, one of the 18 chosen voting systems from one participant of the experiment will be selected randomly to determine the payments.

Just before determining participants' preferences, outcome, and earnings (in this order), all participants of the experiment will be randomly distributed over the groups, so that each group has exactly the number of members for the selected situation. In some situations, the number of participants is larger than the sum of members of the various groups. In this case, some participants will be unlucky – they will not be part of any group and also not receive any payment for part 1 (if one of your choices is implemented you will always be in one of the groups). **It is important to note that when making your decisions you do not yet know which group you will be in later when the payments are determined.**

The 18 decision situations are split in 6 blocks. Groups differ from block to block (3 blocks with 4 groups and 3 blocks with 4 groups). The complexity of the voting systems varies, the "largest" one consist of 7 rows with 5 groups, the "smallest" one of 3 rows with 4 groups.

[Next](#)

Figure 5.17: Instructions choices and payoffs 2, treatments *BI* and *BA*

It is randomly determined whether you prefer X or Y. How is this determined?

- The probability of you preferring either option is one half, independently of all other outcomes.
- Everyone in the same group as you will have the same preference as you. The probability of X being the preference of everyone in the group is one half.
- All participants of this experiment will have the same preference as you. The probability of Y being this preference is one half.

To determine the outcome the groups will vote. This is done after participants have been distributed over the groups. How is it determined whether a group votes for X or for Y?

- Some group members can prefer X while some others prefer Y. If more members of the group prefer X the group votes for X, if more members prefer Y the group votes for Y.
- All members of one group always prefer the same. If the members prefer X the group votes for X, if the members prefer Y the group votes for Y.
- There will be a separate voting stage in which all members determine the group vote, which can be voting for X, for Y, or abstaining.

For showing up and answering all questions carefully you will receive a payment of 12 euros. The earnings from part 1 and 2 of the experiment come on top of this. What about the earnings from part 1 of the experiment?

- You can earn any number of points between 0 and 1800, depending on luck and the different choices you make.
- You will earn 1000 points from part 1 if the outcome is the same as your preference, nothing otherwise.

Send

Figure 5.18: Test questions choices and payoffs



There are 18 different decision situations with two voting systems each. You are asked to always choose the one that you prefer. Afterwards, one situation and one participant are selected and participants are randomly allocated to groups for the selected situation. Next, it is randomly determined for each participant whether he/she prefers X or Y. If the majority in a group prefers X, the group votes for X. If the majority prefers Y, the group votes for Y. Then, the voting system chosen by the selected individual is implemented, which determines whether the outcome is X or Y.

If the outcome is the same as your preference you earn 1000 points, otherwise you earn nothing.

If anything is unclear raise your hand and someone will come to your desk.

[Start experiment](#)

Figure 5.19: Summary screen, treatments *BI* and *FI*

5.A.2 Instruction Differences in the In Front of the Veil Treatments

Figure 5.20 shows a screenshot of the part of the instructions where the treatments *FI* and *FA*, i.e. the in front of the veil treatments, both differ from the instructions of treatment *BI*.



During part 1, you will be asked to make 18 decisions. Each decision consists of choosing one out of two voting systems. You will see a screen with one voting system on the left and one on the right. You choose one of these voting systems by clicking on it. You will see an arrow above each voting system that indicates which group you will be in for payment if your choice is implemented. In the end of the experiment, one of the 18 chosen voting systems from one participant of the experiment will be selected randomly to determine the payments.

Just before determining participants' preferences, outcome, and earnings (in this order), all participants of the experiment will be randomly distributed over the groups, so that each group has exactly the number of members for the selected situation. In some situations, the number of participants is larger than the sum of members of the various groups. In this case, some participants will be unlucky – they will not be part of any group and also not receive any payment for part 1. If one of your choices is implemented you will be in the group that was indicated in the respective choice situation by an arrow.

The 18 decision situations are split in 6 blocks. Groups differ from block to block (3 blocks with 4 groups and 3 blocks with 4 groups). The complexity of the voting systems varies, the "largest" one consist of 7 rows with 5 groups, the "smallest" one of 3 rows with 4 groups.

[Next](#)

Figure 5.20: Instructions choices and payoffs 2, treatments *FI*, and *FA*

5.A.3 Instruction Differences in the Group Aligned Preference Treatments

Figures 5.21 and 5.22 show screenshots of the part of the instructions where the treatments BA and FA , i.e. the group aligned treatments, both differ from the instructions of treatment BI . Note that also the answers to the test questions are partly different, while the questions themselves are not different.



The groups vote to choose between the two possible outcomes X and Y . You know now how the outcome is determined given the votes of the groups and given a voting system (the voting system used will depend on your choices). What the groups vote for depends on the preferences of the group members, which are the participants of the experiment.

Each participant either prefers X or Y . Whether you prefer X or Y will not be determined until the end of the experiment. In each group either all members prefer X or all members prefer Y . The probability that the members of a group prefer X (or Y , similarly) is one half. This probability is the same for all groups independently of the random preferences of participants in other groups.

If the members of a group prefer X the group will vote for X , if the members prefer Y the group will vote for Y . This is done automatically, you cannot change the way your group votes.

The outcome (X or Y , determined by the group voting) will determine your earnings. Your earnings are 1000 points if your preference is equal to the outcome. If not, your earnings are 0.

[Next](#)

Figure 5.21: Instructions choices and payoffs 1, treatments BA and FA



There are 18 different decision situations with two voting systems each. You are asked to always choose the one that you prefer. Afterwards, one situation and one participant are selected and participants are randomly allocated to groups for the selected situation. Next, it is randomly determined for each group whether its members prefer X or Y. If they prefer X, the group votes for X. If they prefer Y, the group votes for Y. Then, the voting system chosen by the selected individual is implemented, which determines whether the outcome is X or Y.

If the outcome is the same as your preference you earn 1000 points, otherwise you earn nothing.

If anything is unclear raise your hand and someone will come to your desk.

[Start experiment](#)

Figure 5.22: Summary screen, treatments BA and FA

Appendix 5.B Properties of the Decision Blocks, Additional Information on the Selection Procedure, and Additional Graphs and Data

5.B.1 Properties of the Decision Blocks

Tables 5.5 to 5.10 contain the properties of the decision blocks. Other than the groups and their sizes and the sets of winning coalitions according to each rule (including the competitor), the tables also show the optimal distribution of the Banzhaf index (PB) or the Shapley-Shubik (SS) index, according to the theoretical Rules I and II, respectively. Furthermore the actual Banzhaf and Shapley-Shubik indices of all the three sets of winning coalitions used are shown. Then the probabilities of being successful as any member (behind the veil treatments) or as a member of a certain group (in front of the veil treatments, ordered from smallest to largest group) for each of the three voting systems are shown, first for independent voting outcome preferences (treatments *BI* and *FI*), then for group aligned voting outcome preferences (treatments *BA* and *FA*). These probabilities have been simulated with two million runs of the voting situation for each of the values attained. Furthermore, the tables show for each of the three voting systems the ‘error terms’ when solving the inverse power problem. Method 1 refers to $err_{group,imp}$ and Method 2 to cv (see Chapter 4). This is done using the Banzhaf index when calculating the error term and the coefficient of variation with respect to the theoretical Rule I and using the Shapley-Shubik index when using the methods with respect to the theoretical Rule II.

Table 5.5: Properties, decision block 1

Constellation of Groups	D:3 C:3 B:15 A:19
Optimal PB index Rule 1	0.1411873 0.1411873 0.3370064 0.3806190
Optimal SS index Rule 2	0.075 0.075 0.375 0.475
Concept Rule 1	{D, C, A}, {D, B, A}, {C, B, A}
Concept Rule 2	{B, A}, {D, B, A}, {C, B, A}
Concept Competitor (c)	{D, C, B}, {D, B, A}, {C, B, A}
Normalized PB index R1	0.2 0.2 0.2 0.4
Normalized PB index R2	0.0 0.0 0.5 0.5
Normalized PB index c	0.2 0.2 0.4 0.2
SS index R1	0.1666667 0.1666667 0.1666667 0.5000000
SS index R2	0.0 0.0 0.5 0.5
SS index c	0.1666667 0.1666667 0.5000000 0.1666667
Success prob, indep pref, behind veil, R1/R2/c	0.5412222 / 0.5416947 / 0.5400489
Success prob, indep pref, per group R1	0.5623765 / 0.562324 / 0.526179 / 0.5464264
Success prob, indep pref, per group R2	0.4997673 / 0.4999598 / 0.5523717 / 0.5464754
Success prob, indep pref, per group c	0.5624157 / 0.5624712 / 0.5523567 / 0.5232603
Success prob, aligned pref, behind veil, R1/R2/c	0.6843514 / 0.7127205 / 0.6719335
Success prob, aligned pref, per group R1	0.6250865 / 0.625031 / 0.625184 / 0.7497865
Success prob, aligned pref, per group R2	0.5002325 / 0.500177 / 0.750038 / 0.7503695
Success prob, aligned pref, per group c	0.625271 / 0.6252155 / 0.7496395 / 0.625331
Error term according R1 (method 1, PB) R1/R2/c	0.08795622 / 0.140436 / 0.1322987
Error term according R2 (method 1, SS) R1/R2/c	0.1335415 / 0.083666 / 0.2286433
Coeff var according R1 (method 2, PB) R1/R2/c	0.3119119 / 0.4252265 / 0.4087207
Coeff var according R2 (method 2, SS) R1/R2/c	0.5840639 / 0.4392977 / 0.6825591

Notes: 'Optimal PB index Rule 1' is the theoretically optimal Banzhaf index per group according to Rule 1, 'Optimal SS index Rule 2' is the theoretically optimal Shapley-Shubik index per group according to Rule 2. The concepts according to Rule 1 and 2 are the sets of winning coalitions performing best in terms of these rules, 'Concept Competitor' is the competing voting system. The normalized PB indices and the SS indices are shown for all three voting systems. Also shown is the probability of success of an individual, behind or in front of the veil of ignorance (in front split up according to groups, ordered from smallest group (here D, can also be E) to largest group (A)). The error terms are the error terms of the inverse power problem $err_{group,imp}$ (Method 1) and cv (Method 2).

Table 5.6: Properties, decision block 2

Constellation of Groups	D:3 C:5 B:7 A:21
Optimal PB index Rule 1	0.1476872 0.1969163 0.2362996 0.4190968
Optimal SS index Rule 2	0.08333333 0.13888889 0.19444444 0.58333333
Concept Rule 1	{B, A}, {D, C, B}, {D, C, A}, {D, B, A}, {C, B, A}
Concept Rule 2	{C, A}, {B, A}, {D, C, A}, {D, B, A}, {C, B, A}
Concept Competitor (c)	{D, A}, {C, A}, {D, C, A}, {D, B, A}, {C, B, A}
Normalized PB index R1	0.1666667 0.1666667 0.3333333 0.3333333
Normalized PB index R2	0.0 0.2 0.2 0.6
Normalized PB index c	0.2 0.2 0.0 0.6
SS index R1	0.1666667 0.1666667 0.3333333 0.3333333
SS index R2	0.000000 0.1666667 0.1666667 0.6666667
SS index c	0.1666667 0.1666667 0.000000 0.6666667
Success prob, indep pref, behind veil, R1/R2/c	0.5526175 / 0.5526517 / 0.5502265
Success prob, indep pref, per group R1	0.5622135 / 0.5469948 / 0.5780369 / 0.5441122
Success prob, indep pref, per group R2	0.4999022 / 0.5469484 / 0.5390204 / 0.566089
Success prob, indep pref, per group comp	0.5623315 / 0.5470398 / 0.5000138 / 0.5659935
Success prob, aligned pref, behind veil, R1/R2/c	0.7221737 / 0.7604319 / 0.7466079
Success prob, aligned pref, per group R1	0.6249915 / 0.625061 / 0.7501065 / 0.749868
Success prob, aligned pref, per group R2	0.499844 / 0.6251955 / 0.624959 / 0.8750155
Success prob, aligned pref, per group comp	0.624664 / 0.6250945 / 0.500139 / 0.8751165
Error term according R1 (method 1, PB) R1/R2/c	0.07923726 / 0.145483 / 0.1737147
Error term according R2 (method 1, SS) R1/R2/c	0.2022253 / 0.06990587 / 0.1099476
Coeff var according R1 (method 2, PB) R1/R2/c	0.2564921 / 0.3686197 / 0.5076884
Coeff var according R2 (method 2, SS) R1/R2/c	0.5433582 / 0.3236694 / 0.5433582

Table 5.7: Properties, decision block 3

Constellation of Groups	D:3 C:5 B:9 A:27
Optimal PB index Rule 1	0.1353527 0.1804702 0.2475020 0.43666750
Optimal SS index Rule 2	0.06818182 0.11363636 0.20454545 0.61363636
Concept Rule 1	{C, A}, {B, A}, {D, C, B}, {D, C, A}, {D, B, A}, {C, B, A}
Concept Rule 2	{D, A}, {C, A}, {B, A}, {D, C, A}, {D, B, A}, {C, B, A}
Concept Competitor (c)	{D, A}, {C, A}, {D, C, B}, {D, C, A}, {D, B, A}, {C, B, A}
Normalized PB index R1	0.08333333 0.25000000 0.25000000 0.41666667
Normalized PB index R2	0.1 0.1 0.1 0.7
Normalized PB index c	0.25000000 0.25000000 0.08333333 0.41666667
SS index R1	0.08333333 0.25000000 0.25000000 0.41666667
SS index R2	0.08333333 0.08333333 0.08333333 0.75000000
SS index c	0.25000000 0.25000000 0.08333333 0.41666667
Success prob, indep pref, behind veil, R1/R2/c	0.5503359 / 0.549906 / 0.5476233
Success prob, indep pref, per group R1	0.5314262 / 0.5704411 / 0.5513113 / 0.5483886
Success prob, indep pref, per group R2	0.5313193 / 0.5235484 / 0.5172294 / 0.5677444
Success prob, indep pref, per group comp	0.593859 / 0.5706518 / 0.5171714 / 0.5483721
Success prob, aligned pref, behind veil, R1/R2/c	0.7558185 / 0.792744 / 0.738921
Success prob, aligned pref, per group R1	0.562579 / 0.687716 / 0.6871145 / 0.8128025
Success prob, aligned pref, per group R2	0.5624065 / 0.5626975 / 0.562096 / 0.937821
Success prob, aligned pref, per group comp	0.6872255 / 0.6875165 / 0.562468 / 0.813002
Error term according R1 (method 1, PB) R1/R2/c	0.03131772 / 0.2186799 / 0.08487629
Error term according R2 (method 1, SS) R1/R2/c	0.1623534 / 0.1205647 / 0.1765774
Coeff var according R1 (method 2, PB) R1/R2/c	0.1692904 / 0.457098 / 0.4173818
Coeff var according R2 (method 2, SS) R1/R2/c	0.4902338 / 0.3370167 / 0.8851774

Table 5.8: Properties, decision block 4

Constellation of Groups	E:1 D:5 C:11 B:13 A:15
Optimal PB index Rule 1	0.0590425 0.1574467 0.2399187 0.2617295 0.2818626
Optimal SS index Rule 2	0.02222222 0.11111111 0.24444444 0.28888889 0.33333333
Concept Rule 1	{E, D, B, A}, {E, C, B, A}, {D, C, B, A}
Concept Rule 2	{C, B, A}, {E, C, B, A}, {D, C, B, A}
Concept Competitor (c)	{D, C, A}, {E, D, C, A}, {D, C, B, A}
Normalized PB index R1	0.1428571 0.1428571 0.1428571 0.2857143 0.2857143
Normalized PB index R2	0.000000 0.000000 0.3333333 0.3333333 0.3333333
Normalized PB index c	0.000000 0.3333333 0.3333333 0.000000 0.3333333
SS index R1	0.10 0.10 0.10 0.35 0.35
SS index R2	0.000000 0.000000 0.3333333 0.3333333 0.3333333
SS index c	0.000000 0.3333333 0.3333333 0.000000 0.3333333
Success prob, indep pref, behind veil, R1/R2/c	0.5246491 / 0.524443 / 0.5214918
Success prob, indep pref, per group R1	0.5621735 / 0.523247 / 0.5153432 / 0.5283421 / 0.5262385
Success prob, indep pref, per group R2	0.499726 / 0.4997985 / 0.5307395 / 0.5284168 / 0.5262444
Success prob, indep pref, per group comp	0.500018 / 0.5466679 / 0.5307172 / 0.5002553 / 0.5261711
Success prob, aligned pref, behind veil, R1/R2/c	0.6012695 / 0.608331 / 0.5859561
Success prob, aligned pref, per group R1	0.562155 / 0.5623885 / 0.562017 / 0.62517 / 0.624909
Success prob, aligned pref, per group R2	0.4995395 / 0.499773 / 0.6246325 / 0.6253375 / 0.6250765
Success prob, aligned pref, per group comp	0.499832 / 0.6250465 / 0.624441 / 0.500064 / 0.624885
Error term according R1 (method 1, PB) R1/R2/c	0.05151498 / 0.08561067 / 0.1622361
Error term according R2 (method 1, SS) R1/R2/c	0.08012336 / 0.06232795 / 0.1775925
Coeff var according R1 (method 2, PB) R1/R2/c	0.3077011 / 0.3985151 / 0.7274828
Coeff var according R2 (method 2, SS) R1/R2/c	0.6102848 / 0.4153242 / 0.8876254

Table 5.9: Properties, decision block 5

Constellation of Groups	E:5 D:5 C:7 B:15 A:17
Optimal PB index Rule 1	0.1449324 0.1449324 0.1739189 0.2594595 0.2767568
Optimal SS index Rule 2	0.1020408 0.1020408 0.1428571 0.3061224 0.3469388
Concept Rule 1	{E, D, C, A}, {E, D, B, A}, {E, C, B, A}, {D, C, B, A}
Concept Rule 2	{C, B, A}, {E, D, B, A}, {E, C, B, A}, {D, C, B, A}
Concept Competitor (c)	{E, C, B}, {E, D, C, B}, {E, D, C, A}, {E, C, B, A}
Normalized PB index R1	0.1764706 0.1764706 0.1764706 0.1764706 0.2941176
Normalized PB index R2	0.06666667 0.06666667 0.2000000 0.33333333 0.33333333
Normalized PB index c	0.33333333 0.06666667 0.33333333 0.20000000 0.06666667
SS index R1	0.15 0.15 0.15 0.15 0.40
SS index R2	0.0500000 0.0500000 0.1333333 0.3833333 0.3833333
SS index c	0.3833333 0.0500000 0.3833333 0.1333333 0.0500000
Success prob, indep pref, behind veil, R1/R2/c	0.5279604 / 0.5271842 / 0.522236
Success prob, indep pref, per group R1	0.5350678 / 0.5353033 / 0.5291155 / 0.5195876 / 0.5306224
Success prob, indep pref, per group R2	0.5117695 / 0.5119502 / 0.5290851 / 0.5326312 / 0.5306096
Success prob, indep pref, per group comp	0.558684 / 0.5117015 / 0.5486878 / 0.5196336 / 0.5060187
Success prob, aligned pref, behind veil, R1/R2/c	0.6155086 / 0.6217915 / 0.581092
Success prob, aligned pref, per group R1	0.594195 / 0.5942325 / 0.5928295 / 0.5936005 / 0.656704
Success prob, aligned pref, per group R2	0.531504 / 0.5315415 / 0.5925965 / 0.6562915 / 0.656471
Success prob, aligned pref, per group comp	0.6567345 / 0.531697 / 0.655349 / 0.593661 / 0.5317055
Error term according R1 (method 1, PB) R1/R2/c	0.04916097 / 0.06425278 / 0.1557973
Error term according R2 (method 1, SS) R1/R2/c	0.09441924 / 0.05338688 / 0.2373642
Coeff var according R1 (method 2, PB) R1/R2/c	0.2099367 / 0.2916223 / 0.8261977
Coeff var according R2 (method 2, SS) R1/R2/c	0.3649335 / 0.2775044 / 1.248378

Table 5.10: Properties, decision block 6

Constellation of Groups	E:3 D:5 C:7 B:13 A:19
Optimal PB index Rule 1	0.1130502 0.1507336 0.1808803 0.2505701 0.3047658
Optimal SS index Rule 2	0.06382979 0.10638298 0.14893617 0.27659574 0.40425532
Concept Rule 1	{D, B, A}, {C, B, A}, {E, D, C, B}, {E, D, C, A}, {E, D, B, A}, {E, C, B, A}, {D, C, B, A}
Concept Rule 2	{E, B, A}, {D, B, A}, {C, B, A}, {E, D, C, A}, {E, D, B, A}, {E, C, B, A}, {D, C, B, A}
Concept Competitor (c)	{B, A}, {E, B, A}, {D, B, A}, {C, B, A}, {E, D, B, A}, {E, C, B, A}, {D, C, B, A}
Normalized PB index R1	0.09090909 0.18181818 0.18181818 0.27272727 0.27272727
Normalized PB index R2	0.1 0.1 0.1 0.3 0.4
Normalized PB index c	0.0 0.0 0.0 0.5 0.5
SS index R1	0.100000 0.1833333 0.1833333 0.2666667 0.2666667
SS index R2	0.08333333 0.08333333 0.08333333 0.25000000 0.50000000
SS index c	0.0 0.0 0.0 0.5 0.5
Success prob, indep pref, behind veil, R1/R2/c	0.5385124 / 0.5378188 / 0.5343319
Success prob, indep pref, per group R1	0.531323 / 0.5471015 / 0.5389068 / 0.5422752 / 0.5346674
Success prob, indep pref, per group R2	0.5315122 / 0.5236948 / 0.5194424 / 0.5423237 / 0.5462195
Success prob, indep pref, per group comp	0.5001042 / 0.5001818 / 0.4999439 / 0.5564911 / 0.5462308
Success prob, aligned pref, behind veil, R1/R2/c	0.6634563 / 0.6726945 / 0.6700662
Success prob, aligned pref, per group R1	0.5625175 / 0.62493 / 0.6245855 / 0.6875055 / 0.6873985
Success prob, aligned pref, per group R2	0.56222 / 0.5622215 / 0.561877 / 0.687208 / 0.750107
Success prob, aligned pref, per group comp	0.5000495 / 0.500051 / 0.4997065 / 0.7493785 / 0.7501495
Error term according R1 (method 1, PB) R1/R2/c	0.02617178 / 0.07484115 / 0.2017992
Error term according R2 (method 1, SS) R1/R2/c	0.09257286 / 0.06799451 / 0.1492592
Coeff var according R1 (method 2, PB) R1/R2/c	0.1173041 / 0.2767331 / 0.6948053
Coeff var according R2 (method 2, SS) R1/R2/c	0.3623343 / 0.255115 / 0.7226806

5.B.2 Further Information on the Selection of the Decision Situations Used in the Experiment

As discussed above, all voting systems corresponding to the rules in this experiment correspond to this rule according to both ways of solving the inverse power problem, based on $err_{group,imp}$ and based on cv . In order to find suitable decision situations, a computer program evaluated (a subset of) all admissible voting systems for all kinds of possible group compositions. The programming was done in R and C. Note that the optimization procedure can be computationally quite involved, because the number of possible sets of winning coalitions grows fast with the number of groups (for N groups, the set of all coalitions has 2^N elements and the set of all different sets of coalitions has 2^{2^N} elements). The programming was done in a way that all computations can be performed on a simple notebook.

Only groups with an odd number of members were considered and the groups cannot be too large in order to still fit into the laboratory. the number of winning coalitions was kept constant across comparisons, as explained in Section 5.3.5. All different decision situations were selected in a way that the recommended rules are not only the same according to both methods of solving the inverse power problem and different from each other, but also in a way that each system does not do too well vis-à-vis the other rule. Furthermore, a ‘competitor’, i.e. a voting system not prescribed by any reasonable normative rule, was added that does not do well according to either rule. In more detail, the selection procedure was done as follows. For each fixed combination of number of groups, number of members per group, and number of winning coalitions the terms that are needed to select the voting system according to both rules and both methods to solve the inverse power were calculated. Next, all situations were dismissed where the two different methods do not yield the same unique outcome (separately for Rule I and Rule II). Then, situations were dismissed where the recommendations of Rule I and Rule II coincide. One also wants these recommendations not to be too similar in terms of ‘performance’ according to the respective other rule. Therefore, only voting systems were considered where the respective error term and the coefficient of variation sufficiently differ. The system recommended by Rule I has at least a 15% higher value of $err_{group,imp}$ and cv than the voting system recommended by Rule II. Vice versa, the Rule II system has similarly higher values of these terms than the Rule I system. The competitor similarly has higher error terms when compared to each of the two rules. 15% might seem a bit arbitrary – it is chosen as high as possible so that it is still possible to have a variety in terms of group constellations, given the constraints on group size to be feasible in the laboratory. More information on the computer programs is available on request.

5.B.3 Additional Graphs and Data

Figure 5.23 shows the proportion of people predominantly choosing one voting system. This graph is similar to Figure 5.3, but it also includes the graphs if one considers a system to be chosen predominantly only if it has been chosen at least 10 or 11 out of 12 times.

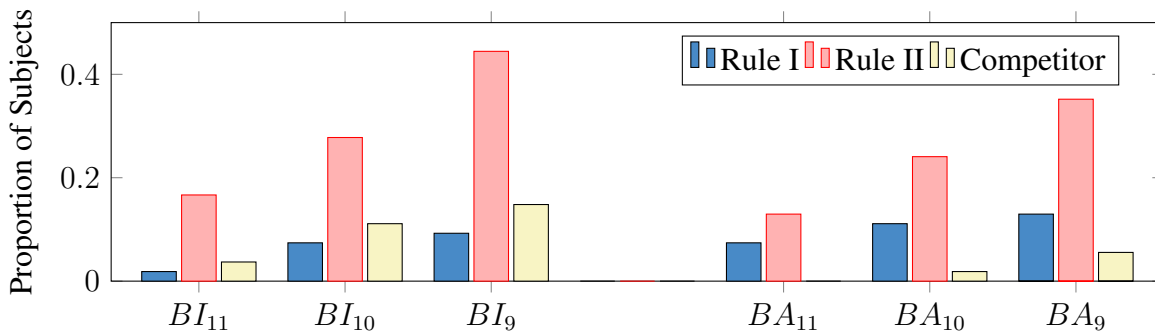


Figure 5.23: Proportion of participants predominantly choosing one system

Notes: The figure shows the proportion of participants that predominantly chose one system. BI_{11} shows how many participants chose a system at least 11 out of 12 possible times in treatment BI , etc.

Figure 5.24 shows the choices of participants in all six decision blocks. Each of the six graphs (a) to (f) represents one block (the block number corresponds to the order of blocks as in Table 5.2, not to the order as the blocks appeared in the experiment, which is random). In each graph, the treatments are shown next to one another. For the in front of the veil treatments, the data has been split according to which group a participant will be in for payment if her choices are selected. Only the choices of the participants of the smallest (FI_S and FA_S) and largest (FI_L and FA_L) groups are depicted here (the numbers of observations then drop to between 10 and 17). The bars show how one voting system was chosen over another: The bar ‘R1-R2’ shows how often the system recommended by Rule I was chosen over the system recommended by Rule II, the bar ‘R1-c’ shows how often the system of Rule I was chosen over the respective additional competitor of the block, similarly for ‘R2-c’. The scale used is difference in proportion, i.e. if the Rule II voting system was chosen 70% of times in comparison with the corresponding Rule I voting system, the value of the corresponding bar ‘R1-R2’ would be -0.4 (the difference between 0.3 and 0.7). The data underlying this graph can be found in Table 5.11.

Table 5.12 shows, split according to decision block and treatment (and group for the in front of the veil treatments), how often one voting system was preferred over both other voting systems.

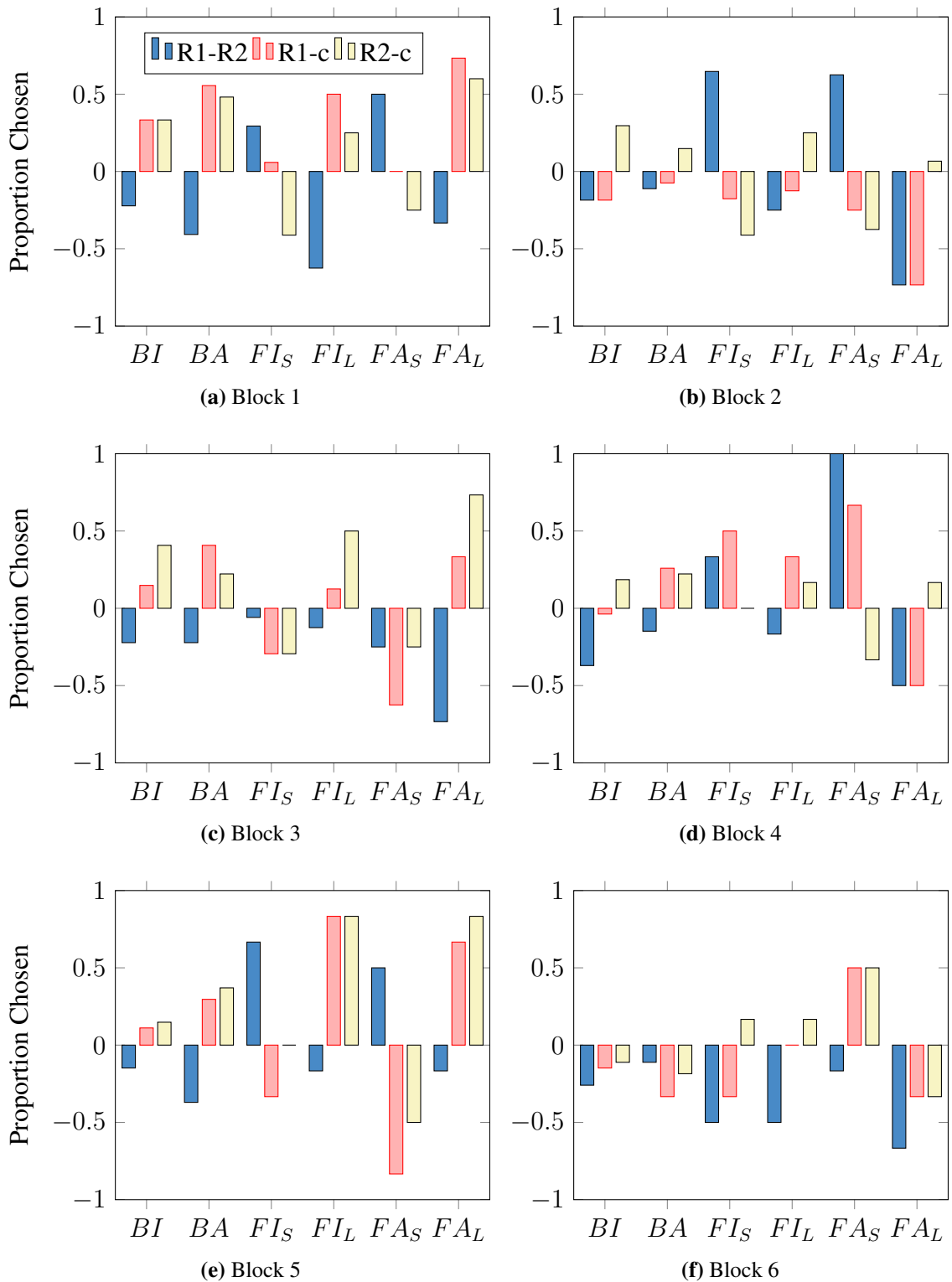


Figure 5.24: Overview of the data

Notes: The sub-figures (a) to (f) show the choices of participants in each decision block. The bars ‘R1-R2’ show how often voting systems according to Rule I have been chosen over voting systems according to Rule II (similar for ‘R1-c’ and ‘R2-c’). These choices are shown for all subjects in treatments BI and BA and for the subjects of the smallest and largest groups in treatments FI and FA (FI_S , FI_L , FA_S , and FA_L , respectively).

Table 5.11: Choice proportion data

Treat.	Choice	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
<i>BI</i> (54)	R1-R2	0.3888889	0.4074074	0.3888889	0.3148148	0.4259259	0.3703704
	R1-c	0.6666667	0.4074074	0.5740741	0.4814815	0.5555556	0.4259259
	R2-c	0.6666667	0.6481481	0.7037037	0.5925926	0.5740741	0.4444444
<i>BA</i> (54)	R1-R2	0.2962963	0.4444444	0.3888889	0.4259259	0.3148148	0.4444444
	R1-c	0.7777778	0.462963	0.7037037	0.6296296	0.6481481	0.3333333
	R2-c	0.7407407	0.5740741	0.6111111	0.6111111	0.6851852	0.4074074
<i>FI_A</i> (16/12)	R1-R2	0.1875	0.375	0.4375	0.4166667	0.4166667	0.25
	R1-c	0.75	0.4375	0.5625	0.6666667	0.9166667	0.5
	R2-c	0.625	0.625	0.75	0.5833333	0.9166667	0.5833333
<i>FI_B</i> (11/11)	R1-R2	0.1818182	0.8181818	0.7272727	0.09090909	0.2727273	0.4545455
	R1-c	0.2727273	0.6363636	0.9090909	0.72727273	0.6363636	0.4545455
	R2-c	0.6363636	0.8181818	0.8181818	0.90909091	0.8181818	0.4545455
<i>FI_C</i> (14/11)	R1-R2	0.7142857	0.5714286	0.71428571	0	0.3636364	1
	R1-c	0.7142857	0.4285714	0.42857143	0	0.2727273	1
	R2-c	0.2142857	0.6428571	0.07142857	0.6363636	0.2727273	0.9090909
<i>FI_D</i> (17/12)	R1-R2	0.6470588	0.8235294	0.4705882	0.58333333	0.8333333	0.9166667
	R1-c	0.5294118	0.4117647	0.3529412	0.08333333	0.8333333	0.9166667
	R2-c	0.2941176	0.2941176	0.3529412	0.25	0.5833333	0.75
<i>FI_E</i> (0/12)	R1-R2	NaN	NaN	NaN	0.6666667	0.8333333	0.25
	R1-c	NaN	NaN	NaN	0.75	0.3333333	0.3333333
	R2-c	NaN	NaN	NaN	0.5	0.5	0.5833333
<i>FA_A</i> (15/12)	R1-R2	0.3333333	0.1333333	0.1333333	0.25	0.4166667	0.1666667
	R1-c	0.8666667	0.1333333	0.6666667	0.25	0.8333333	0.3333333
	R2-c	0.8	0.5333333	0.8666667	0.5833333	0.9166667	0.3333333
<i>FA_B</i> (11/11)	R1-R2	0.18181818	0.9090909	0.9090909	0.3636364	0	0.36363636
	R1-c	0.09090909	1	0.9090909	0.8181818	0.1818182	0.09090909
	R2-c	0.63636364	1	0.7272727	0.7272727	0.6363636	0.27272727
<i>FA_C</i> (14/11)	R1-R2	0.7142857	0.4285714	0.6428571	0.1818182	0.3636364	0.8181818
	R1-c	0.7142857	0.5	0.5	0.1818182	0.2727273	0.8181818
	R2-c	0.2857143	0.7142857	0.2857143	0.4545455	0.3636364	0.9090909
<i>FA_D</i> (16/10)	R1-R2	0.75	0.8125	0.375	1	0.6	0.5
	R1-c	0.5	0.375	0.1875	0.5	0.8	0.8
	R2-c	0.375	0.3125	0.375	0.4	0.7	0.8
<i>FA_E</i> (0/12)	R1-R2	NaN	NaN	NaN	1	0.75	0.4166667
	R1-c	NaN	NaN	NaN	0.8333333	0.08333333	0.75
	R2-c	NaN	NaN	NaN	0.3333333	0.25	0.75

Notes: This table shows the choices that the participants made, split according to treatment and block, as illustrated in Figure 5.24. Numbers in parentheses are the numbers of observations; where split, the first number refers to the first three blocks and the second number to the last three blocks.

Table 5.12: Counts of the most preferred voting system per block

Treatment	System	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
<i>BI</i> (54)	Rule I	12	15	8	11	9	13
	Rule II	28	22	27	23	22	11
	competitor	9	11	10	19	19	20
<i>BA</i> (54)	Rule I	11	19	13	17	13	15
	Rule II	34	16	27	21	27	5
	competitor	7	11	8	10	9	27
<i>FI_A</i> (16/12)	Rule I	2	5	5	4	5	2
	Rule II	9	5	9	3	6	5
	competitor	3	5	2	3	0	3
<i>FI_B</i> (11/11)	Rule I	1	7	7	1	1	2
	Rule II	5	2	3	9	8	4
	competitor	3	2	0	1	1	5
<i>FI_C</i> (14/11)	Rule I	6	6	4	0	1	11
	Rule II	1	4	1	7	2	0
	competitor	4	2	7	4	7	0
<i>FI_D</i> (17/12)	Rule I	5	6	3	0	9	10
	Rule II	5	2	4	2	2	0
	competitor	7	8	10	9	1	1
<i>FI_E</i> (0/12)	Rule I	NaN	NaN	NaN	7	3	3
	Rule II	NaN	NaN	NaN	3	2	4
	competitor	NaN	NaN	NaN	0	6	5
<i>FA_A</i> (15/12)	Rule I	4	1	1	1	4	2
	Rule II	8	8	11	6	6	3
	competitor	1	6	1	4	1	7
<i>FI_B</i> (11/11)	Rule I	1	10	10	4	0	1
	Rule II	6	1	0	6	7	3
	competitor	4	0	1	1	4	7
<i>FI_C</i> (14/11)	Rule I	6	5	4	1	0	7
	Rule II	3	7	3	5	3	2
	competitor	4	1	6	4	6	1
<i>FI_D</i> (16/10)	Rule I	5	5	1	5	5	5
	Rule II	4	2	4	0	3	3
	competitor	6	7	8	5	0	1
<i>FI_E</i> (0/12)	Rule I	NaN	NaN	NaN	10	1	4
	Rule II	NaN	NaN	NaN	0	2	5
	competitor	NaN	NaN	NaN	2	8	0

Notes: For each treatment condition and decision block, this table shows how many participant preferred each system over the other two. Numbers in parentheses are the numbers of observations; where split, the first number refers to the first three blocks and the second number to the last three blocks.

Bibliography

- Adam, K. (2007). Experimental Evidence on the Persistence of Output and Inflation. *The Economic Journal*, 117(520):603–636.
- Aleskerov, F., Belianin, A., and Pogorelskiy, K. (2009). Power and preferences: an experimental approach. *Available at SSRN 1574777*.
- Alon, N. and Edelman, P. H. (2010). The inverse Banzhaf problem. *Social Choice and Welfare*, 34(3):371–377.
- Anand, P. and Van Hees, M. (2006). Capabilities and achievements: An empirical study. *The Journal of Socio-Economics*, 35(2):268–284.
- Andreoni, J. (1990). Impure altruism and donations to public goods: a theory of warm-glow giving. *The Economic Journal*, 100(401):464–477.
- Anufriev, M., Assenza, T., Hommes, C. H., and Massaro, D. (2013). Interest rate rules and macroeconomic stability under heterogeneous expectations. *Macroeconomic Dynamics*, 17:1574–1604.
- Anufriev, M. and Hommes, C. H. (2012). Evolutionary selection of individual expectations and aggregate outcomes. *American Economic Journal: Microeconomics*, 4:35–64.
- Arifovic, J. and Sargent, T. (2003). Laboratory experiments with an expectational phillips curve. In *Evolution and Procedures in Central Banking*, pages 23–55. Cambridge University Press.
- Arthur, W. B., Holland, J. H., LeBaron, B., Palmer, R. G., and Tayler, P. (1997). Asset pricing under endogenous expectations in an artificial stock market. In Arthur, W. B., Durlauf, S., and Lane, D., editors, *The Economy as an Evolving Complex System II*. Addison-Wesley.
- Assenza, T., Bao, T., Hommes, C., and Massaro, D. (2014a). Experiments on expectations in macroeconomics and finance. In Duffy, J., editor, *Experiments in Macroeconomics*, volume 17 of *Research in Experimental Economics*.

- Assenza, T., Heemeijer, P., Hommes, C., and Massaro, D. (2014b). Managing self-organization of expectations through monetary policy: a macro experiment. *CeNDEF Working Paper 14-07*.
- Atkinson, A. and Stiglitz, J. (1980). *Lectures on public economics*. Economics handbook series. McGraw-Hill Book Co.
- Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263.
- Baldwin, R. and Widgrén, M. (2004). Winners and losers under various dual majority rules for the EU Council of Ministers. *CEPR Discussion Paper No. 4450*.
- Balinski, M. L. and Young, H. P. (2001). *Fair representation: Meeting the ideal of one man, one vote*. Brookings Institution Press, 2nd edition.
- Banzhaf III, J. F. (1964). Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19:317.
- Bao, M., Hommes, C., Sonnemans, J., and Tuinstra, J. (2012). Individual expectations, limited rationality and aggregate outcomes. *Journal of Economic Dynamics and Control*, 36:1101–1120.
- Barbera, S. and Jackson, M. O. (2006). On the weights of nations: Assigning voting weights in a heterogeneous union. *Journal of Political Economy*, 114(2):317–339.
- Barthélémy, F. and Martin, M. (2011). A comparison between the methods of apportionment using power indices: the case of the us presidential elections. *Annals of Economics and Statistics*, 101/102:87–106.
- Bassi, M. (2010). Mirrlees meets laibson: Optimal income taxation with bounded rationality. *CSEF Working Paper*.
- Becker, G. M., Degroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Systems Research and Behavioral Science*, 9:226–232.
- Beisbart, C., Bovens, L., and Hartmann, S. (2005). A utilitarian assessment of alternative decision rules in the Council of Ministers. *European Union Politics*, 6(4):395–418.
- Benhabib, J., Evans, G. W., and Honkapohja, S. (2014). Liquidity traps and expectation dynamics: Fiscal stimulus or fiscal austerity? *Journal of Economic Dynamics and Control*, 45:220–238.

- Bernasconi, M. and Kirchkamp, O. (2000). Why do monetary policies matter? an experimental study of saving and inflation in an overlapping generations model. *Journal of Monetary Economics*, 46(2):315–343.
- Blais, A., Laslier, J.-F., Poinas, F., and Van der Straeten, K. (2014). Why voters like voting rules: Self-interest, ideology, or sincerity? *Working Paper*.
- Bloomfield, R. and Hales, J. (2002). Predicting the next step of a random walk: Experimental evidence of regime-switching beliefs. *Journal of Financial Economics*, 65:397–414.
- Blumkin, T., Ruffle, B. J., and Ganun, Y. (2012). Are income and consumption taxes ever really equivalent? Evidence from a real-effort experiment with real goods. *European Economic Review*, 56(6):1200 – 1219.
- Bolton, G. E., Brandts, J., and Ockenfels, A. (2005). Fair procedures: Evidence from games involving lotteries. *The Economic Journal*, 115(506):1054–1076.
- Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.
- Branch, W. (2004). The Theory of Rationally Heterogeneous Expectations: Evidence from Survey Data on Inflation Expectations. *The Economic Journal*, 114(497):592–621.
- Branch, W. and McGough, B. (2009). A new keynesian model with heterogeneous expectations. *Journal of Economic Dynamics and Control*, 33:1036–1051.
- Branch, W. and McGough, B. (2010). Dynamic predictors selection in a new keynesian model with heterogeneous expectations. *Journal of Economic Dynamics and Control*, 34(8):1492–1508.
- Branch, W. A., Carlson, J., Evans, G. W., and McGough, B. (2009). Monetary policy, endogenous inattention and the volatility trade-off. *The Economic Journal*, 119(534):123–157.
- Brock, W. A. and Hommes, C. H. (1997). A rational route to randomness. *Econometrica*, 65(5):1059–1095.
- Brock, W. A. and Hommes, C. H. (1998). Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic Dynamics and Control*, 22:1235–1274.
- Bullard, J. and Mitra, K. (2002). Learning about monetary policy rules. *Journal of Monetary Economics*, 49(6):1105–1129.

- Camerer, C. F., Loewenstein, G., and Rabin, M. (2011). *Advances in behavioral economics*. Princeton University Press.
- Carroll, C. (2003). Macroeconomic expectations of households and professional forecasters. *Quarterly Journal of Economics*, 118:269–298.
- Charness, G. and Kuhn, P. (2011). Lab labor: What can labor economists learn from the lab? In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 4, pages 229 – 330. Elsevier.
- Charness, G., Rigotti, L., and Rustichini, A. (2007). Individual behavior and group membership. *American Economic Review*, 97(4):1340–1352.
- Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, 99(4):1145–77.
- Clarida, R., Gali, J., and Gertler, M. (2000). Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory. *Quarterly Journal of Economics*, 115(1):147–180.
- Coleman, J. S. (1971). Control of collectivities and the power of a collectivity to act. *Social Choice*, pages 269–300.
- Conlisk, J. (1996). Why bounded rationality? *Journal of Economic Literature*, 34:669–700.
- Cornand, C. and M'Baye, C. K. (2013). Does inflation targeting matter? An experimental investigation. *GATE Working Paper 30*.
- Cowell, F. (2011). *Measuring inequality*. Oxford University Press, 3rd edition.
- Cox, J. C., Rider, M., and Sen, A. (2012). Tax incidence: Do institutions matter? An experimental study. *Experimental Economics Center Working Paper 2012-17*.
- Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–74.
- Davis, D. and Korenok, O. (2011). Nominal shocks in monopolistically competitive markets: An experiment. *Journal of Monetary Economics*, 58(6):578–589.
- De, A., Diakonikolas, I., and Servedio, R. (2012). The inverse Shapley value problem. In Czumaj, A., Mehlhorn, K., Pitts, A., and Wattenhofer, R., editors, *Automata, Languages, and Programming*, pages 266–277. Springer.
- De Bartolome, C. A. M. (1995). Which tax rate do people use: Average or marginal? *Journal of Public Economics*, 56(1):79–96.

- De Cremer, D. and Tyler, T. R. (2007). The effects of trust in authority and procedural fairness on cooperation. *Journal of Applied Psychology*, 92(3):639.
- De Grauwe, P. (2011). Animal spirits and monetary policy. *Economic Theory*, 47(2-3):423–457.
- De Grauwe, P. (2012a). Booms and busts in economic activity: A behavioral explanation. *Journal of Economic Behavior & Organization*, 83(3):484–501.
- De Grauwe, P. (2012b). *Lectures on behavioral macroeconomics*. Princeton University Press.
- De Grauwe, P. and Kaltwasser, P. R. (2012). Animal spirits in the foreign exchange market. *Journal of Economic Dynamics and Control*, 36(8):1176–1192.
- De Nijs, F. and Wilmer, D. (2012). Evaluation and improvement of Laruelle-Widgrén inverse Banzhaf approximation. *arXiv preprint arXiv:1206.1145*.
- Diks, C. and van der Weide, R. (2005). Herding, a-synchronous updating and heterogeneity in memory in a cbs. *Journal of Economic Dynamics and Control*, 29:741–763.
- Djanali, I. and Sheehan-Connor, D. (2012). Tax affinity hypothesis: Do we really hate paying taxes? *Journal of Economic Psychology*, 33(4):758–775.
- Drouvelis, M., Montero, M., and Sefton, M. (2010). Gaining power through enlargement: Strategic foundations and experimental evidence. *Games and Economic Behavior*, 69(2):274–292.
- Duffy, J. (2012). Macroeconomics: A survey of laboratory research. *University of Pittsburgh Working Paper*.
- Esposito, G., Guerci, E., Hanaki, N., Lu, X., and Watanabe, N. (2012). An experimental study on “meaningful learning” in weighted voting games. *Working Paper*.
- Evans, G. W. and Honkapohja, S. (2001). *Learning and Expectations in Macroeconomics*. Princeton University Press.
- Evers, M., De Mooij, R., and Van Vuuren, D. (2008). The wage elasticity of labour supply: A synthesis of empirical estimates. *De Economist*, 156:25–43.
- Falk, A. and Fehr, E. (2003). Why labour market experiments? *Labour Economics*, 10(4):399–406.

- Falk, A. and Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, 326(5952):535–538.
- Fatima, S., Wooldridge, M., and Jennings, N. R. (2008). An anytime approximation method for the inverse Shapley value problem. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, volume 2, pages 935–942.
- Fehr, E. and Goette, L. (2007). Do workers work more if wages are high? Evidence from a randomized field experiment. *American Economic Review*, 97(1):298–317.
- Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.
- Felsenthal, D. S. and Machover, M. (2004). Analysis of QM rules in the draft constitution for Europe proposed by the European Convention, 2003. *Social Choice and Welfare*, 23(1):1–20.
- Ferrer-i Carbonell, A. (2013). Happiness economics. *SERIEs*, pages 1–26.
- Finkelstein, A. (2009). E-ztax: Tax salience and tax rates. *The Quarterly Journal of Economics*, 124(3):969–1010.
- Fochmann, M., Kiesewetter, D., Blaufus, K., Hundsdoerfer, J., and Weimann, J. (2010a). Tax Perception – An Empirical Survey. *Arqus Discussion Paper No. 99*.
- Fochmann, M., Kiesewetter, D., and Sadrieh, A. (2012). Investment behavior and the biased perception of limited loss deduction in income taxation. *Journal of Economic Behavior and Organization*, 81(1):230 – 242.
- Fochmann, M. and Weimann, J. (2013). The effects of tax salience and tax experience on individual work efforts in a framed field experiment. *FinanzArchiv/Public Finance Analysis*, 69:511–542.
- Fochmann, M., Weimann, J., Blaufus, K., Hundsdoerfer, J., and Kiesewetter, D. (2010b). Grosswage illusion in a real effort experiment. *FEMM Working Paper*.
- Frankel, J. and Froot, K. (1987). Using survey data to test standard propositions regarding exchange rate expectations. *American Economic Review*, 77(1):133–153.
- Frankel, J. and Froot, K. (1991). Chartists, fundamentalists and the demand for dollars. In Courakis, A. and Taylor, M., editors, *Private Behaviour and Government Policy in Interdependent Economies*. Oxford: Clarendon.

- Galí, J. (2008). *Monetary Policy, Inflation and the Business Cycle: An Introduction to the New Keynesian Framework*. Princeton University Press.
- Geller, C., Mustard, J., and Shahwan, R. (2012). An experimental investigation of pure voting power. In *Selected papers from the 33rd Australian Conference of Economists, Sydney Sept. 27-30, 2004*, pages 1–21. Blackwell.
- Gigerenzer, G. and Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT Press.
- Gigerenzer, G. and Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press.
- Giritligil, A. E. and Sertel, M. R. (2005). Does majoritarian approval matter in selecting a social choice rule? An exploratory panel study. *Social Choice and Welfare*, 25(1):43–73.
- Grether, D. and Plott, C. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review*, 69(4):623–638.
- Guerci, E., Hanaki, N., Watanabe, N., Esposito, G., and Lu, X. (2014). A methodological note on a weighted voting experiment. *Social Choice and Welfare*, 43(4):827–850.
- Guesnerie, R. (2009). Macroeconomic and Monetary Policies from the Eductive Viewpoint. In *Monetary Policy under Uncertainty and Learning*, volume 13 of *Central Banking, Analysis, and Economic Policies Book Series*, chapter 6, pages 171–202. Central Bank of Chile.
- Hayashi, A., Nakamura, B. K., and Gamage, D. (2013). Experimental evidence of tax salience and the labor-leisure decision: Anchoring, tax aversion, or complexity? *Public Finance Review*, 41(2):203–226.
- Heemeijer, P., Hommes, C. H., Sonnemans, J., and Tuinstra, J. (2009). Price stability and volatility in markets with positive and negative expectations feedbacks: An experimental investigation. *Journal of Economic Dynamics and Control*, 33:1052–1072.
- Holm, P., Honkapohja, S., and Koskela, E. (1994). A monopoly-union model of wage determination with capital and taxes: An empirical application to the Finnish manufacturing. *European Economic Review*, 38(2):285–303.
- Hommes, C., Massaro, D., and Weber, M. (2015). Monetary policy under behavioral expectations: Theory and experiment. *Tinbergen Institute Discussion Paper 15-087/II*.
- Hommes, C. H. (2011). The heterogeneous expectations hypothesis: Some evidence from the lab. *Journal of Economic Dynamics and Control*, 35:1–24.

- Hommes, C. H., Huang, H., and Wang, D. (2005a). A robust rational route to randomness in a simple asset pricing model. *Journal of Economic Dynamics and Control*, 29:1043–1072.
- Hommes, C. H., Sonnemans, J., Tuinstra, J., and van de Velden, H. (2005b). Coordination of expectations in asset pricing experiments. *Review of Financial Studies*, 18(3):955–980.
- Hommes, C. H., Sonnemans, J., Tuinstra, J., and van de Velden, H. (2008). Expectations and bubbles in asset pricing experiments. *Journal of Economic Behavior and Organization*, 67(1):116–133.
- Hossain, T. and List, J. A. (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, 58(12):2151–2167.
- Huck, S., Normann, H., and Oechssler, J. (2004). Two are few and four are many: number effects in experimental oligopolies. *Journal of Economic Behavior and Organization*, 53:435–446.
- Irtel, H. (2007). Pxlabs: The psychological experiments laboratory [online]. version 2.1.11. mannheim (germany): University of mannheim. [cited 25 april 2011]. available from <http://www.pxlab.de>.
- Kahneman, D., Knetsch, J., and Tversky, A. (1991). Anomalies: The endowment effect, loss aversion and status quo bias. *Journal of Economic Perspectives*, 5(1):193–206.
- Kahneman, D. and Krueger, A. B. (2006). Developments in the measurement of subjective well-being. *Journal of Economic Perspectives*, 20(1):3–24.
- Kaniovski, S. (2008). The exact bias of the Banzhaf measure of power when votes are neither equiprobable nor independent. *Social Choice and Welfare*, 31(2):281–300.
- Kelley, H. and Friedman, D. (2002). Learning to forecast price. *Economic Inquiry*, 40(4):556–573.
- Kirsch, W. and Langner, J. (2011). Invariably suboptimal: An attempt to improve the voting rules of the Treaties of Nice and Lisbon. *Journal of Common Market Studies*, 49(6):1317–1338.
- Koriyama, Y., Macé, A., Treibich, R., and Laslier, J.-F. (2013). Optimal apportionment. *Journal of Political Economy*, 121(3):584–608.
- Koskela, E. and Schöb, R. (1999). Does the composition of wage and payroll taxes matter under nash bargaining? *Economics Letters*, 64(3):343–349.

- Krawczyk, M. and Le Lec, F. (2010). 'Give me a chance!' An experiment in social decision under risk. *Experimental Economics*, 13(4):500–511.
- Kryvtsov, O. and Petersen, L. (2013). Expectations and monetary policy: Experimental evidence. *Bank of Canada Working Paper 13-44*.
- Kurz, M. (2011). A new keynesian model with diverse beliefs. *Stanford University Working Paper*.
- Kurz, M., Piccillo, G., and Wu, H. (2013a). Modeling diverse expectations in an aggregated new keynesian model. *Journal of Economic Dynamics and Control*, 37:1403–1433.
- Kurz, S. (2012). On the inverse power index problem. *Optimization*, 61(8):989–1011.
- Kurz, S., Maaser, N., and Napel, S. (2013b). On the democratic weights of nations. *Working Paper*.
- Kurz, S., Maaser, N., Napel, S., and Weber, M. (2015). Mostly sunny: A forecast of tomorrow's power index research. *Homo Oeconomicus*, 32(1):133–146.
- Kurz, S. and Napel, S. (2014). Heuristic and exact solutions to the inverse power index problem for small voting bodies. *Annals of Operations Research*, 215(1):137–163.
- Lang, P. J. (1985). *The cognitive psychophysiology of emotion: Anxiety and anxiety disorders*. Lawrence Erlbaum.
- Laruelle, A. and Valenciano, F. (2002). Inequality among EU citizens in the EU's Council decision procedure. *European Journal of Political Economy*, 18(3):475–498.
- Laruelle, A. and Valenciano, F. (2008). *Voting and collective decision-making: Bargaining and power*. Cambridge University Press.
- Laruelle, A. and Valenciano, F. (2010). Egalitarianism and utilitarianism in committees of representatives. *Social Choice and Welfare*, 35(2):221–243.
- Le Breton, M., Montero, M., and Zaporozhets, V. (2012). Voting power in the EU Council of Ministers and fair decision making in distributive politics. *Mathematical Social Sciences*, 63(2):159–173.
- Le Breton, M. and Van Der Straeten, K. (2014). Influence vs. utility in the evaluation of voting rules: A new look at the penrose formula. *Toulouse School of Economics Working Paper TSE-511*.

- Leech, D. (2002). Designing the voting system for the Council of the European Union. *Public Choice*, 113(3-4):437–464.
- Leech, D. (2003). Power indices as an aid to institutional design: The generalised apportionment problem. *Jahrbuch für Neue Politische Ökonomie*, 22:107–121.
- Lehmann, E., Marical, F., and Rioux, L. (2013). Labor income responds differently to income-tax and payroll-tax reforms. *Journal of Public Economics*, 99:66–84.
- Lei, V. and Noussair, C. N. (2002). An experimental test of an optimal growth model. *American Economic Review*, 92(3):549–570.
- Maaser, N. and Napel, S. (2007). Equal representation in two-tier voting systems. *Social Choice and Welfare*, 28(3):401–420.
- Mankiw, N., Reis, R., and Wolfers, J. (2003). Disagreement about inflation expectations. In Gertler, M. and Rogoff, K., editors, *NBER Macroeconomics Annual*.
- Manski, C. and McFadden, D. (1981). Alternative estimators and sample designs for discrete choice analysis. In Manski, C. and McFadden, D., editors, *Structural Analysis of Discrete Data*. MIT Press.
- Marcet, A. and Nicolini, J. P. (2003). Recurrent hyperinflations and learning. *The American Economic Review*, 93(5):1476–1498.
- Marimon, R. and Sunder, S. (1993). Indeterminacy of equilibria in a hyperinflationary world: Experimental evidence. *Econometrica*, 61:1073–1107.
- Massaro, D. (2013). Heterogeneous expectations in monetary dsge models. *Journal of Economic Dynamics and Control*, 37(3):680–692.
- Matějka, F. and McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298.
- McCaffery, E. and Slemrod, J. (2006). *Behavioral public finance*. Russell Sage.
- Montero, M., Sefton, M., and Zhang, P. (2008). Enlargement and the balance of power: an experimental study. *Social Choice and Welfare*, 30(1):69–87.
- Morton, R. B. (1987). A group majority voting model of public good provision. *Social Choice and Welfare*, 4(2):117–131.
- Mullainathan, S., Schwartzstein, J., and Congdon, W. J. (2012). A reduced-form approach to behavioral public finance. *Annual Review of Economics*, 4(1):511–540.

- Muysken, J., Van Veen, T., and De Regt, E. (1999). Does a shift in the tax burden create employment? *Applied Economics*, 31(10):1195–1205.
- Napel, S. and Widgrén, M. (2006). The inter-institutional distribution of power in EU codecision. *Social Choice and Welfare*, 27(1):129–154.
- Oswald, A. J. (1985). The economic theory of trade unions: An introductory survey. *The Scandinavian Journal of Economics*, 87(2):160–193.
- Penrose, L. (1946). The elementary statistics of majority voting. *Journal of the Royal Statistical Society*, 109(1):53–57.
- Pfajfar, D. and Zakelj, B. (2014). Experimental evidence on inflation expectations formation. *Journal of Economic Dynamics and Control*, 44:147–168.
- Picard, P. and Toulemonde, E. (2001). On the equivalence of taxes paid by employers and employees. *Scottish Journal of Political Economy*, 48(4):461–470.
- Pissarides, C. A. (2000). *Equilibrium unemployment theory*. MIT Press.
- Rasmussen, B. (1997). Non-equivalence of employment and payroll taxes in imperfectly competitive labor markets. *Working Paper*.
- Riedl, A. and Tyran, J.-R. (2005). Tax liability side equivalence in gift-exchange labor markets. *Journal of Public Economics*, 89(11-12):2369–2382.
- Robins, P. K. (1985). A comparison of the labor supply findings from the four negative income tax experiments. *Journal of Human Resources*, 20(4):567–582.
- Rupert, T. J. and Wright, A. M. (1998). The use of marginal tax rates in decision making: The impact of tax rate visibility. *Journal of the American Taxation Association*, 20(2):83–99.
- Saez, E., Matsaganis, M., and Tsakloglou, P. (2012). Earnings determination and taxes: Evidence from a cohort-based payroll tax reform in Greece. *The Quarterly Journal of Economics*, 127(1):493–533.
- Sausgruber, R. and Tyran, J.-R. (2005). Testing the mill hypothesis of fiscal illusion. *Public Choice*, 122(1-2):39–68.
- Sausgruber, R. and Tyran, J.-R. (2011). Are we taxing ourselves?: How deliberation and experience shape voting on taxes. *Journal of Public Economics*, 95(1-2):164–176.
- Schokkaert, E. (2007). Capabilities and satisfaction with life. *Journal of Human Development*, 8(3):415–430.

- Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, 12(2):225–237.
- Schram, A. J. (1990). A dynamic model of voter behavior and the demand for public goods among social groups in Great Britain. *Journal of Public Economics*, 41(2):147–182.
- Sertel, M. R. and Giritligil, A. E. (2003). Selecting a social choice rule: An exploratory panel study. In Sertel, M. R. and Koray, S., editors, *Advances in Economic Design*, Springer.
- Shapiro, C. and Stiglitz, J. E. (1984). Equilibrium unemployment as a worker discipline device. *The American Economic Review*, 74(3):433–444.
- Shapley, L. and Shubik, M. (1954). A method for evaluating the distribution of power in a committee system. *American Political Science Review*, 48(03):787–792.
- Shapley, L. S. (1953). A value for n-person games. *The Annals of Mathematical Statistics*, (28):307–317.
- Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118.
- Stiglitz, J. (2000). *Economics of the public sector*. Norton.
- Sugden, R. (1984). Reciprocity: the supply of public goods through voluntary contributions. *The Economic Journal*, 94:772–787.
- Sutter, M. (2000). Fair allocation and re-weighting of votes and voting power in the EU before and after the next enlargement. *Journal of Theoretical Politics*, 12(4):433–449.
- Sutter, M. (2009). Individual behavior and group membership: Comment. *American Economic Review*, 99(5):2247–2257.
- Turnovec, F. (2009). Fairness and squareness: Fair decision making rules in the EU council? *IES Working Paper No. 1/2009*.
- Turnovec, F. (2011). Fair voting rules in committees, strict proportional power and optimal quota. *Homo Oeconomicus*, 27(4):463–479.
- Turnovec, F., Mercik, J., and Mazurkiewicz, M. (2008). Power indices methodology: decisiveness, pivots and swings. In Braham, M. and Steffen, F., editors, *Power, Freedom and Voting, Essays in Honor of Manfred J. Holler*, pages 23–37.
- Tversky, A. and Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185:1124–1130.

- Tversky, A. and Thaler, R. (1990). Anomalies: Preference reversal. *Journal of Economic Perspectives*, 4(2):201–211.
- Ullmann, R. and Watrin, C. (2008). Comparing direct and indirect taxation: The influence of framing on tax compliance. *European Journal of Comparative Economics*, 5(1):23–56.
- Van Huyck, J. B., Cook, J. P., and Battalio, R. C. (1994). Selection dynamics, asymptotic stability, and adaptive behavior. *Journal of Political Economy*, 102(5):975–1005.
- Weber, M. (2009). Simultaneous estimation of covariate coefficients and connection signs. *Diplomarbeit, Faculty of Mathematics and Physics, University of Freiburg*.
- Weber, M. (2011). Income tax or employer’s contribution – an experiment on the perception of ‘economically equivalent’ duties. *Tinbergen Institute Mphil-Thesis*.
- Weber, M. (2015a). Choosing the rules: Preferences over voting systems in assemblies of representatives. *CREED Working Paper*.
- Weber, M. (2015b). Two-tier voting: Measuring inequality and specifying the inverse power problem. *CREED Working Paper*.
- Weber, M. and Schram, A. (2015). The non-equivalence of labor market taxes: A real-effort experiment. *CREED Working Paper*.
- Weber, M., Schumacher, M., and Binder, H. (2014). Regularized regression incorporating network information: Simultaneous estimation of covariate coefficients and connection signs. *Tinbergen Institute Discussion Paper 14-089/I*.
- Woodford, M. (2003). *Interest and prices: Foundations of a theory of monetary policy*. Princeton University Press.
- Woodford, M. (2010). Robustly optimal monetary policy with near-rational expectations. *American Economic Review*, 100(1):274–303.
- Woodford, M. (2013). Macroeconomic analysis without the rational expectations hypothesis. *Annual Review of Economics*, 5(1):303–346.
- Życzkowski, K. and Słomczyński, W. (2013). Square root voting system, optimal threshold and π . In Holler, M. J. and Nurmi, H., editors, *Power, Voting, and Voting Power: 30 Years After*, pages 573–592. Springer.

Summary

This thesis consists of four essays dealing with topics relevant for the public sector. The essays cover issues from different parts of economics, partly overlapping with political science. They reach from labor taxation over monetary policy to the preference for voting institutions. Throughout this thesis it is, in contrast to classical economics, not assumed that humans are necessarily fully rational and selfish. Once full rationality is no longer assumed, experiments become an important tool to learn about human behavior and to test the hypotheses arising from behavioral theories. Therefore, most of the work in this thesis makes use of experimental methods.

The first essay (Chapter 2) deals with labor market taxation. A classic economic result is that under full rationality a labor market tax levied on employers and a corresponding income tax levied on employees are equivalent. With boundedly rational agents, this equivalence is no longer obvious and the different reactions to these two taxes become important for policy making, political economics, and optimal taxation theory. This chapter studies the differential effects of the two taxes on preferences concerning the size of the public sector, subjective well-being, labor supply, and on-the-job performance. This is done in a real-effort laboratory experiment. The findings suggest that employer-side taxes induce preferences for a larger public sector. In addition, subjective well-being is higher when the taxes are levied on employers while labor supply is lower, at least at the extensive margin. The essay also discusses three mechanisms that may underlie these results. These mechanisms are based on (i) considering a euro of wage more salient than a euro of tax, (ii) considering a tax a loss which looms larger than the plain effect of having lower net earnings, and (iii) experiencing pleasure when other people benefit from one's own tax payments (more than when other people benefit in the same way from other people's tax payments).

The second essay (Chapter 3) is concerned with aggregate macroeconomic behavior and its implications for monetary policy. Expectations play a crucial role in modern macroeconomic models. In this chapter, the common assumption of rational expectations in a New Keynesian framework is replaced by the assumption that expectations are formed according to a heuristics switching model. This behavioral model of expectation formation assumes

that individuals make their expectations using relatively simple heuristics. These heuristics are based on the past behavior of the relevant variables and individuals decide which heuristics they use based on how well these heuristics have performed in earlier forecasts. This chapter studies how the economy behaves under the rational and the behavioral models of expectation formation with a special focus on price stability, more precisely on inflation volatility. Contrary to the rational model, the behavioral model predicts that inflation volatility can be lowered if a central bank reacts to the output gap in addition to inflation. These opposing theoretical predictions are then tested in a learning to forecast experiment in the laboratory. The only difference between the treatments lies in the parameters of the monetary policy equation simulating the behavior of the central bank. The experimental results support the behavioral model and the claim that reacting to the output gap in addition to inflation can indeed lower inflation volatility.

The two last essays are concerned with voting institutions. The focus in these essays lies on situations in which different groups make collective decisions by voting in an assembly with each group represented by a single person. Such voting takes place in a wide variety of institutions, including the Council of the European Union, UN General Assembly, German Bundesrat, ECB, and thousands of boards of directors and professional and non-professional associations.

The third essay (Chapter 4) is the only one in this thesis not making use of experimental methods. It assumes throughout that a variable of interest is given standing for a citizen's power (or influence or representation) in light of a voting system. This variable could be an expected payoff, but it could also be a more behavioral variable such as for example the probability of influencing the voting outcome, which can be measured in different ways. This chapter is about two closely related problems. The first problem is how to measure the inequality of a voting system. The second is called the inverse power problem: the problem of finding voting systems that approximate a voting power distribution as well as possible. The most common normative theoretical rules on the design of voting systems prescribe that indirect voting power should be as equal as possible for all individuals. This chapter argues that the coefficient of variation is an appropriate measure of inequality of a voting system and the appropriate way to specify the inverse power problem. For the inverse power problem, using the coefficient of variation turns out to be equivalent to an error term minimizing the distance of a normalized indirect voting power vector from the fair ideal. Furthermore, it is illustrated in this chapter that using objective functions that only consider (weighted) voting power at the group level to specify the inverse problem is suboptimal.

Also the fourth essay (Chapter 5) is concerned with situations in which groups make collective decisions by voting in an assembly where each group is represented by a single person. Although there is a vast theoretical, normative literature on the question of what

voting system such an assembly should use, to date there is no consensus. Instead of studying properties of voting systems based on theoretical concepts, this essay studies which voting systems individuals actually prefer. This is important for the legitimacy and acceptance of voting institutions. It can furthermore have an influence on peoples' behavior as people react to the institutions and procedures in place. It is investigated in a laboratory experiment which voting systems participants choose behind the veil of ignorance, that is, when they do not know which group they will be in. As a control, also participants' choices in front of the veil are observed, that is, when they do know which group they will be in. Behind the veil of ignorance, participants predominantly choose voting systems that allocate more voting power to larger groups than the most prominent theoretical concept (called Penrose's Square Root Rule) suggests. Participants choose voting systems much more often that have the property that voting power as measured by the Shapley-Shubik power index is proportional to group size. Furthermore, participants in front of the veil of ignorance behave differently than behind the veil of ignorance. When participants know which group they will be in they choose to the benefit of their own group.

Samenvatting (Summary in Dutch)

Deze proefschrift is opgebouwd uit vier essays, die allen relevant zijn voor de publieke sector. De essays behandelen onderwerpen die vallen binnen verschillende gebieden van de economische wetenschappen en deels overlappen met de politicologie. Ze reiken van belasting op arbeid tot monetair beleid en de voorkeur voor steminstituties. In tegenstelling tot wat gebruikelijk is in de neoklassieke economie wordt in dit proefschrift niet aangenomen dat mensen in alle omstandigheden volledig rationeel en egoïstisch zijn. Zodra volledige rationaliteit niet langer wordt aangenomen, worden experimenten een belangrijk gereedschap om meer te weten te komen over menselijk gedrag en om de hypothesen die voortkomen uit gedragstheorieën te onderzoeken. Bij het merendeel van het werk in deze thesis wordt daarom gebruik gemaakt van experimentele methodes.

Het eerste essay (Hoofdstuk 2) behandelt belastingen op de arbeidsmarkt. Een klassiek economisch resultaat is dat in het geval van volledige rationaliteit een werkgeversbelasting en de corresponderende inkomstenbelasting (opgelegd aan werknemers) equivalent zijn. Dat komt omdat de effecten van de belasting via marktwerking op de arbeidsmarkt worden doorgespeeld naar de lonen. In het geval van beperkt rationele agenten is deze equivalentie echter niet langer vanzelfsprekend en worden verschillen in de reacties op deze twee belastingen belangrijk voor beleidsvorming, politieke economie en optimale belastingtheorie. Dit hoofdstuk bestudeert de verschillende effecten van de twee belastingen op (i) de voorkeuren met betrekking tot de grootte van de publieke sector, (ii) subjectief welzijn, en (iii) arbeidsaanbod en werkprestaties. Dit wordt gedaan met behulp van een laboratorium-experiment waarin deelnemers taken uitvoeren tegen een beloning. De resultaten suggereren dat belastingen aan de werkgeverszijde voorkeuren teweegbrengen voor een grotere publieke sector. Bovendien is het subjectieve welzijn groter bij werkgeversbelastingen terwijl het arbeidsaanbod lager is. Het essay beschrijft verder drie mechanismen die aan deze resultaten ten grondslag kunnen liggen. Deze mechanismen zijn gebaseerd op (i) het idee dat een euro salaris meer opvalt dan een euro belasting, (ii) het idee dat een betaalde belasting als een verlies wordt beschouwd, dat verder reikt dan enkel het feit dat men een lager netto inkomen heeft en (iii) het ervaren van plezier wanneer andere mensen voordeel hebben van de eigen

belastingafdracht (meer dan wanneer andere mensen op dezelfde wijze voordeel hebben van andermans belastingafdracht).

Het tweede essay (Hoofdstuk 3) draait om geaggregeerd macro-economisch gedrag en de gevolgen daarvan voor het monetaire beleid. Uitgangspunt is dat verwachtingen een cruciale rol spelen in moderne macro-economische modellen. In dit hoofdstuk wordt de vaak gedane aanname van rationele verwachtingen in een Nieuw Keynesiaans kader vervangen door de aanname dat verwachtingen worden gevormd volgens een zogeheten ‘heuristisch wisselmodel’. Dit gedragsmodel van verwachttingsvorming gaat uit van het feit dat individuen hun verwachtingen bepalen op basis van relatief eenvoudige heuristieken (vuistregels). Individueel beslissen welke vuistregel ze gebruiken op basis van hoe goed deze heeft gefunctioneerd bij eerdere voorspellingen. Dit hoofdstuk bestudeert hoe de economie zich gedraagt op basis van een gedragsmodel dat bouwt op het gebruik van heuristieken en vergelijkt dit met het rationele model van verwachttingsvorming. De focus is hierbij op prijsstabiliteit, of meer precies, op inflatievolatiliteit. In tegenstelling tot het rationele model voorspelt het gedragsmodel dat inflatievolatiliteit kan worden verlaagd als een centrale bank niet alleen kijkt naar inflatie, maar ook reageert op de ontwikkeling van de arbeidsmarkt (meer precies op de zogenaamde ‘output gap’). Deze tegengestelde theoretische voorspellingen worden vervolgens getest in een ‘learning-to-forecast’ experiment in het laboratorium. Het enige verschil tussen de experimentele condities zit in de parameters van de monetaire beleidsvergelijking die het gedrag van de centrale bank simuleert. De experimentele resultaten ondersteunen het gedragsmodel en de claim dat het reageren op de ‘output gap’ in aanvulling op inflatie inderdaad de inflatievolatiliteit kan verlagen.

De laatste twee essays behandelen steminstituten. De focus in deze essays ligt op situaties waarin verschillende groepen collectieve beslissingen nemen door te stemmen in een vergadering, waarbij elke groep wordt vertegenwoordigd door één enkele persoon. Dit soort institutie komt in de praktijk veel voor, bijvoorbeeld in de Raad van de Europese Unie, de Algemene Vergadering van de VN, de Duitse Bondsraad, het ECB en duizenden besturen en professionele en niet-professionele verenigingen.

Het derde essay (Hoofdstuk 4) is het enige in dit proefschrift waarin geen gebruik wordt gemaakt van experimenten. Het uitgangspunt hier is dat een variabele wordt ontwikkeld, die de invloed van een burger in het kader van een stelsysteem weerspiegelt. Deze variabele kan gebaseerd zijn op een verwachte uitkomst. Het zou ook een meer op gedrag gerichte variabele kunnen zijn, zoals bijvoorbeeld de kans op beïnvloeding van de stemuitslag, hetgeen op verschillende manieren gemeten kan worden. Dit hoofdstuk draait vervolgens om twee sterk gerelateerde problemen. Het eerste probleem is hoe men de ongelijkheid van een stelsysteem kan meten. Het tweede wordt het ‘inverse power’ probleem genoemd: het vinden van stelsystemen die een bepaalde verdeling van invloed zo goed mogelijk benaderen. De

meest gebruikte normatieve theoretische regels voor het ontwerp van stelsystemen schrijven voor dat indirecte stemmacht zo gelijk mogelijk verdeeld dient te zijn over alle individuen. In dit hoofdstuk wordt gesteld dat de variatiecoëfficiënt een geschikte maat voor de ongelijkheid van een stelsysteem is en een adequate manier biedt om het inverse power probleem te specificeren. Bovendien wordt in dit hoofdstuk geïllustreerd dat het gebruik van doel-functies die alleen (gewogen) stemmacht op groepsniveau gebruiken om het inverse power probleem te specificeren suboptimaal is.

Ook in het vierde essay (Hoofdstuk 5) wordt ingegaan op situaties waarin groepen collectieve beslissingen nemen door te stemmen in een vergadering, waarbij elke groep één vertegenwoordiger heeft. Hoewel er een grote hoeveelheid theoretische, normatieve literatuur beschikbaar is met betrekking tot de vraag welk stelsysteem bij een dergelijke bijeenkomst kan of moet worden gebruikt, is daar tot op heden nog geen consensus over. In plaats van het bestuderen van de eigenschappen van stelsystemen gebaseerd op theoretische concepten, bestudeert dit essay de vraag aan welke stelsystemen individuen eigenlijk de voorkeur geven. Dit is van belang voor de legitimiteit en acceptatie van steminstituten. Het kan bovendien van invloed zijn op het gedrag van mensen, omdat mensen reageren op de instituten en procedures die er zijn. Middels een laboratoriumexperiment wordt onderzocht welke stelsystemen participanten kiezen in een situatie waarin zij nog niet weten in welke groep zij zullen zitten. In een dergelijke situatie van onwetendheid kiezen participanten voornamelijk voor stelsystemen waarbij meer stemmacht wordt toegekend aan grotere groepen dan het meest prominente theoretische concept (dat de Wet van Penrose wordt genoemd) stelt. Participanten kiezen veel vaker voor stelsystemen die de eigenschap hebben dat stemmacht, zoals gemeten door de Shapley-Shubik machtsindex, proportioneel is aan groeps grootte. Daarnaast gedragen participanten zich in een situatie van onwetendheid anders dan in een situatie van kennis. Wanneer participanten weten in welke groep zij zullen zitten, kiezen ze in het voordeel van hun eigen groep.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

583. L.T. GATAREK, *Econometric Contributions to Financial Trading, Hedging and Risk Measurement*
584. X. LI, *Temporary Price Deviation, Limited Attention and Information Acquisition in the Stock Market*
585. Y. DAI, *Efficiency in Corporate Takeovers*
586. S.L. VAN DER STER, *Approximate feasibility in real-time scheduling: Speeding up in order to meet deadlines*
587. A. SELIM, *An Examination of Uncertainty from a Psychological and Economic Viewpoint*
588. B.Z. YUESHEN, *Frictions in Modern Financial Markets and the Implications for Market Quality*
589. D. VAN DOLDER, *Game Shows, Gambles, and Economic Behavior*
590. S.P. CEYHAN, *Essays on Bayesian Analysis of Time Varying Economic Patterns*
591. S. RENES, *Never the Single Measure*
592. D.L. IN 'T VELD, *Complex Systems in Financial Economics: Applications to Interbank and Stock Markets*
593. Y. YANG, *Laboratory Tests of Theories of Strategic Interaction*
594. M.P. WOJTOWICZ, *Pricing Credits Derivatives and Credit Securitization*
595. R.S. SAYAG, *Communication and Learning in Decision Making*
596. S.L. BLAUW, *Well-to-do or doing well? Empirical studies of wellbeing and development*
597. T.A. MAKAREWICZ, *Learning to Forecast: Genetic Algorithms and Experiments*
598. P. ROBALO, *Understanding Political Behavior: Essays in Experimental Political Economy*
599. R. ZOUTENBIER, *Work Motivation and Incentives in the Public Sector*
600. M.B.W. KOBUS, *Economic Studies on Public Facility use*
601. R.J.D. POTTER VAN LOON, *Modeling non-standard financial decision making*
602. G. MESTERS, *Essays on Nonlinear Panel Time Series Models*
603. S. GUBINS, *Information Technologies and Travel*
604. D. KOPÁNYI, *Bounded Rationality and Learning in Market Competition*

605. N. MARTYNOVA, *Incentives and Regulation in Banking*
606. D. KARSTANJE, *Unraveling Dimensions: Commodity Futures Curves and Equity Liquidity*
607. T.C.A.P. GOSENS, *The Value of Recreational Areas in Urban Regions*
608. Ł.M. MARĆ, *The Impact of Aid on Total Government Expenditures*
609. C. LI, *Hitchhiking on the Road of Decision Making under Uncertainty*
610. L. ROSENDAHL HUBER, *Entrepreneurship, Teams and Sustainability: a Series of Field Experiments*
611. X. YANG, *Essays on High Frequency Financial Econometrics*
612. A.H. VAN DER WEIJDE, *The Industrial Organization of Transport Markets: Modeling pricing, Investment and Regulation in Rail and Road Networks*
613. H.E. SILVA MONTALVA, *Airport Pricing Policies: Airline Conduct, Price Discrimination, Dynamic Congestion and Network Effects.*
614. C. DIETZ, *Hierarchies, Communication and Restricted Cooperation in Cooperative Games*
615. M.A. ZOICAN, *Financial System Architecture and Intermediation Quality*
616. G. ZHU, *Three Essays in Empirical Corporate Finance*
617. M. PLEUS, *Implementations of Tests on the Exogeneity of Selected Variables and their Performance in Practice*
618. B. VAN LEEUWEN, *Cooperation, Networks and Emotions: Three Essays in Behavioral Economics*
619. A.G. KOPÁNYI-PEUKER, *Endogeneity Matters: Essays on Cooperation and Coordination*
620. X. WANG, *Time Varying Risk Premium and Limited Participation in Financial Markets*
621. L.A. GORNICKA, *Regulating Financial Markets: Costs and Trade-offs*
622. A. KAMM, *Political Actors playing games: Theory and Experiments*
623. S. VAN DEN HAUWE, *Topics in Applied Macroeconometrics*
624. F.U. BRÄUNING, *Interbank Lending Relationships, Financial Crises and Monetary Policy*
625. J.J. DE VRIES, *Estimation of Alonso's Theory of Movements for Commuting*
626. M. POPLAWSKA, *Essays on Insurance and Health Economics*
627. X. CAI, *Essays in Labor and Product Market Search*
628. L. ZHAO, *Making Real Options Credible: Incomplete Markets, Dynamics, and Model Ambiguity*
629. K. BEL, *Multivariate Extensions to Discrete Choice Modeling*
630. Y. ZENG, *Topics in Trans-boundary River sharing Problems and Economic Theory*